Statistical Methods for Effect Estimation in Biomedical Research: Robustness and Efficiency

Statistical Methods for Effect Estimation in Biomedical Research: Robustness and Efficiency

A thesis presented

by

Matthew Steven Cefalu

to

The Department of Biostatistics

in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the subject of
Biostatistics

Harvard University
Cambridge, Massachusetts

May 2013

Dissertation Advisor: Professor Francesca Dominici Matthew Steven Cefalu

Statistical Methods for Effect Estimation in Biomedical Research: Robustness and Efficiency

Abstract

Practical application of statistics in biomedical research is predicated on the notion that one can readily return valid effect estimates of the health consequences of treatments (exposures) that are being studied. The goal as statisticians should be to provide results that are scientifically useful, to use the available data as efficiently as possible, to avoid unnecessary assumptions, and, if necessary, develop methods that are robust to incorrect assumptions. In this dissertation, I provide methods for effect estimation that meet these goals. I consider three scenarios: (1) clustered binary outcomes; (2) continuous outcomes with a binary treatment; and (3) continuous outcomes with potentially missing continuous exposure. In each of these settings, I discuss the shortfalls of current statistical methods for effect estimation available in the literature and propose new and innovative methods that meet the previously stated goals. The validity of each proposed estimator is theoretically verified using asymptotic arguments, and the finite sample behavior is studied through simulation.

Contents

List of Figures

List of Tables

To my loving companions, Kristen and Rigley.

Acknowledgments

I would like to take this time to briefly thank those who have been most influential in my academic career, and apologize to those who I miss. First, to my parents and brother Michael, thank you for your continuous love and support. To Kristen Barger, none of this would have been possible without you. To my advisor, Francesca Dominici, thank you for the outstanding mentorship. I would also like to acknowledge Eric Tchetgen Tchetgen and Giovanni Parmigiani for their contributions to this dissertation and for the knowledge that they shared with me. I thank May Boggess for her support early in my statistical career, who enlightened me to the possibility of continuing my education beyond an undergraduate program. Finally, I thank all of my friends and extended family.

1. Introduction

Practical application of statistics in biomedical research is predicated on the notion that one can readily return valid effect estimates of the health consequences of treatments (exposures) that are being studied. The goal as statisticians should be to provide results that are scientifically useful, to use the available data as efficiently as possible, to avoid unnecessary assumptions, and, if necessary, develop methods that are robust to incorrect assumptions.

In randomized clinical trials, where control over the treatment assignment is possible, comparing the effectiveness of the treatments is a fairly straightforward endeavor. One relies on the random treatment assignment to ensure that the treatment groups are balanced with regards to covariates that influence the outcome. However, in situations where a randomized clinical trial is not feasible, researchers rely on epidemiological evidence to estimate the effect of different treatments.

For example, consider the problem of estimating the effect of air pollution on cardiovascular health. A randomized clinical trial designed to answer this question would randomly assign individual to receive differing doses of air pollution, and require the participants to receive the assigned level of air pollution for a prolonged period of time. Such a study is not feasible, as it is not ethical to expose individuals to an exposure (air pollution) that is known to have detrimental health effects. Additionally, it is not clear how one would deliver the necessary dose of air pollution without locking the participant in a chamber that exposes them to a constant level of pollution.

As an alternative, one can consider the epidemiological evidence that air pollution adversely effects cardiovascular health. In such a setting, the spatiotemporal variation in air pollution and the cardiovascular outcome would be used to estimate the association of interest. The use of "association" was by choice, as it will be difficult to make any causal conclusions since the exposure was not randomized. Due to the lack of randomization, there may exist other factors that influence both exposure and outcome on the same spatiotemporal scale (i.e. daily temperature), and as such, will not allow us to prop-

erly estimate the effect of air pollution on cardiovascular health.

This problem is not unique to studies of air pollution and health. Many studies must rely on observational data in which the exposure has not been randomized, and do so to estimate the health effect of interest. The field of causal inference has attempted to address this issue by trying to recreate a hypothetical randomized trial based on the observational data. The potential outcomes framework of Rubin (1974) gives a theoretical foundation defining a causal effect, and subsequent methodological developments use potential outcomes to perform valid causal inference from observational data (see Rosenbaum and Rubin (1983) and Robins et al. (2000) as starting points on relevant literature).

In this dissertation, I address the problem of effect estimation in biomedical research by first defining health effects that are scientifically meaningful. I consider three scenarios: (1) estimating risk ratios from clustered binary outcomes; (2) estimating the average causal effect of a binary treatment on a continuous outcome; and (3) estimating the linear effect of a continuous exposure on a continuous outcomes with missing data in the exposure. For each health effect of interest, I discuss the shortfalls of current statistical methods available in the literature and propose new and innovative methods that meet the previously stated goals of robustness and efficiency with minimal assumptions. The validity of each proposed estimator is theoretically verified using asymptotic arguments, and the finite sample behavior is studied through simulation.

In Chapter 2, I discuss estimating the risk ratio of a treatment or exposure on a binary outcome when there is clustering in the data. Such data could arise from a cluster randomized trial or from a study with repeated measures on an individual (e.g. longitudinal data). In Chapter 3, I discuss estimating the average causal effect of a binary treatment on a continuous outcome. I propose a new class of estimators for the average causal effect, the model averaged double robust estimators, that account for model uncertainty in both the propensity score and outcome model through the use of model averaging. The model averaged double robust estimators extend the desirable double robustness property by

achieving consistency under the much weaker assumption that either the true propensity score model or the true outcome model be within a specified, possibly large, class of models. In Chapter 4, I introduce the concept of bias inflation due to exposure prediction of a confounded effect estimate by simultaneously considering exposure prediction and confounding adjustment. I derive a closed form expression for the bias of an effect estimate when using a predicted exposure that decomposes into the product of two pieces: the bias due to the lack of adjustment for confounding and a bias inflation factor due to predicting the exposure.

2. Efficient estimation of risk ratios from clustered binary data

[1]Matthew Cefalu and [1,2]Eric Tchetgen Tchetgen

[1]Department of Biostatistics, Harvard School of Public Health
[2]Department of Epidemiology, Harvard School of Public Health

Abstract

Risk ratios are often the target of inference in epidemiologic studies. The log-binomial model is a natural choice that readily returns risk ratios, but suffers from well known convergence issues. Alternate methods have been proposed to estimate risk ratios for a common binary outcome; however, there has been little work in estimating risk ratios for clustered binary data. The modified Poisson regression approach can be used to take clustering into account through the use of generalized estimating equations, but leads to a potentially inefficient estimator due to the incorrect distributional assumption. In this article, we derive an estimate of the risk ratio that accounts for clustering in the outcome, does not rely on an estimate of the baseline risk for consistency, and delivers asymptotically efficient estimates of the risk ratio parameter. An alternative efficient estimator is provided that bounds the predicted probability by 1, thus guaranteeing stable performance of the estimator. A simulation study is provided verifying that the proposed estimator outperforms the modified Poisson approach as well as estimators that assume no clustering. We apply our method to the *Young Citizens* study, a cluster randomized trial involving a behavioral intervention deigned to train children aged 10-14 years to educate their communities about HIV.

2.1 Introduction

Risk ratios are often the target of inference in epidemiologic studies. They allow a researcher to easily evaluate the multiplicative association between risk factors and binary outcomes. The log binomial model (Wacholder, 1986) is a natural choice that readily returns risk ratios, but suffers from well known convergence issues (Zou, 2004). The traditional approach to avoid convergence issues is to report odds ratios by using logistic regression as the odds ratio provides a good approximation of the risk ratio when the outcome is rare. However, it is often the case that the outcome is not rare within all levels

of risk factors, and using logistic regression will lead to overestimation of the risk ratio. Further, the odds ratio effect measure may be misinterpreted by non-experts (Knol et al., 2011).

Several methods have been proposed to estimate risk ratios for a common binary outcome (Wacholder, 1986; Lee, 1994; Skove et al., 1998; Greenland, 2004; Zou, 2004; Spiegelman and Hertzmark, 2005; Chu and Cole, 2010; Tchetgen Tchetgen, 2012). Each of these methods, except for Lee (1994) and Tchetgen Tchetgen (2012), share the requirement that the log-baseline risk must be estimated in order to obtain a consistent estimate of the risk ratios. This requirement is not easily satisfied, and may lead to a violation of the model restriction that all predicted probabilities are less than 1. Worse, failure to satisfy the model conditions often results in a lack of convergence of the estimation procedures.

Recently, methods have been proposed to address these issues. Chu and Cole (2010) developed a Bayesian approach that incorporates the model restriction in the estimation procedure, while Tchetgen Tchetgen (2012) presents a frequentist approach that allows for consistent and efficient estimation of the risk ratios that does not rely on obtaining an estimate for the baseline risk. It was shown that a simple plug-in estimate of the baseline risk may be used without altering the large sample efficiency of the estimated risk ratios. Another, the modified Poisson regression approach, has been widely cited and adopted as a simple method of risk ratio estimation for both observational and intervention studies (Zou, 2004). This method uses a Poisson distribution for the data in place of the Bernoulli distribution.

However, there has been little work in estimating risk ratios for clustered binary data. Such data could arise from a cluster randomized trial or from a study with repeated measures on an individual (e.g. longitudinal data). Yelland et al. (2011) provide evidence that the modified Poisson regression approach can be used to take clustering into account through the use of generalized estimating equations (GEE) (Liang and Zeger, 1986). They showed that for both observational and intervention studies, the modified Poisson regres-

sion approach using GEEs to account for clustering results in small relative bias and near nominal confidence interval coverage. A major drawback of this approach is that the covariance structure is guaranteed to be misspecified because of the incorrect distributional assumption, leading to a potentially inefficient estimator. Note that the misspecified covariance structure is by choice and is chosen to improve numerical convergence.

In this article, we generalize the work of Tchetgen Tchetgen (2012) to allow for clustered outcomes in the estimation of risk ratios. We show that our method does not rely on an estimate of the baseline risk for consistency and delivers asymptotically efficient estimates of the risk ratios. A slight modification to the approach is described that guarantees the estimated probabilities are bounded by 1. Therefore, the method guarantees stable performance of the estimated risk ratios. We provide a simulation study under both correct and incorrect specification of the working correlation structure that verifies the proposed estimator outperforms the modified Poisson approach as well as estimators that assume no clustering.

We apply our method to the *Young Citizens* study (Kamo et al., 2008), a cluster randomized trial involving a behavioral intervention deigned to train children aged 10-14 years to educate their communities about HIV.

2.2 Methods

2.2.1 Independent outcomes

To begin, we give a brief review of the work of Tchetgen Tchetgen (2012). Consider independent binary outcomes Y_i and a set of q covariates X_i with:

$$log(P(Y_i = 1|X_i)) = log(\mathrm{E}[Y_i|X_i]) = \alpha_0 + X_i\beta_0$$

where the parameter of interest is the q-dimensional vector of log relative risks, β_0.

Tchetgen Tchetgen (2012) provided a simple estimator of β_0 that is asymptotically efficient, in the sense that it has the minimal variance of any regular and asymptotically linear (Bickel et al., 1998) estimator of β_0. Specifically, a large class of estimators was derived that contains many common estimators of the risk ratio as well as the semiparametric efficient estimator. First, an initial consistent estimate of β_0 is provided that is free of the intercept and can be constructed by solving the equation $0 = \sum_{i:Y_i=1}(Z_i - \exp\{\widehat{\beta}W_i\})W_i$, where $W_i = -(X_i - \bar{X})$ and $Z_i = 0$ for all i. This corresponds to an artificial case only model in which the pseudo-outcome Z_i is assumed to follow a Poisson distribution with mean given by the intercept-free multiplicative model $\exp(\beta W_i)$, which facilitates its use with standard regression software. Then, the class of one-step update estimators is given by:

$$\widehat{\beta}(w) = \widehat{\beta} + \left[\sum_i Y_i \widehat{T}_i(w) X_i^T\right]^{-1} \left[\sum_i Y_i \widehat{T}_i(w)\right]$$

where $\widehat{\beta}$ is an initial consistent estimate of β_0 and

$$\widehat{T}_i(w) = \left\{ w_i - \frac{\sum_i w_i \exp(\widehat{\beta}^T X_i)}{\sum_i \exp(\widehat{\beta}^T X_i)} \right\}$$

It was shown that $w_i = X_i$ is asymptotically equivalent to the Breslow-Lee estimator, $w_i = \exp(-\widehat{\beta}^T X_i)(X_i - \mathbf{X})$ returns $\widehat{\beta}$ exactly, and $\widehat{\beta}(w_{opt})$ is asymptotically efficient, with

$$w_{opt,i} = (1 - \widehat{p}_i)^{-1} \left[X_i - \frac{\sum_i X_i (1 - \widehat{p}_i)^{-1} \widehat{p}_i}{\sum_i (1 - \widehat{p}_i)^{-1} \widehat{p}_i}\right]$$

and

$$\widehat{p}_i = \exp(\widehat{\beta}^T X_i) \sum_j Y_j \exp(-\widehat{\beta}^T X_j)/n$$

In general, the difficulty in estimating β_0 lies in the fact that an estimate of the predicted risk $\widehat{p}_i$ must be provided and must be such that predicted probability is bounded by 1 on the support of X. The estimator $\widehat{\beta}(w_{opt})$ (and hence $\widehat{p}_i$) uses a simple plug-in estimate for the log-baseline risk, but any consistent estimate of α_0 could be used without affecting the large sample efficiency of $\widehat{\beta}(w_{opt})$. However, this does not guarantee the predicted probability is bounded by 1 on the support of X. Tchetgen Tchetgen (2012) provides a solution that bounds the predicted probability without requiring an estimate of the baseline risk and will be discussed in detail in Section 2.3.1

2.2.2 Correlated outcomes

We generalize the approach of Tchetgen Tchetgen (2012) to allow for correlation among the outcomes. Let $\mathbf{Y}_i$ be a k-dimensional response vector and $\mathbf{X}_i$ be a (kxq) matrix of covariates for $i = 1, \ldots, n$. Consider the semiparametric model with the only restriction

$$\mathrm{E}\left[\mathbf{Y}|\mathbf{X}\right] = \mu(\mathbf{X}|\alpha_0, \beta_0) = \exp\left(\alpha_0 \mathbf{1}_k + \mathbf{X}\beta_0\right)$$

where β_0 is a q-dimensional parameter of interest. Note that all observations share a common intercept, but this assumption can easily be relaxed as discussed in Section 2.3.2 below. The key in the derivation of our estimator is that our model is semiparametric in the sense that we allow the intercept and the dependence between outcomes to remain unrestricted by treating them as nuisance parameters. As a result, our inferences are robust to misspecification of the baseline risk and working covariance structure.

We briefly review the principles of semiparametric theory. Consider a model $\mathcal{M}$ with parameters (ϕ, η), where ϕ is a finite dimensional parameter of interest and η is a potentially infinite dimensional nuisance parameter. Define the nuisance tangent space Λ for the semiparametric model $\mathcal{M}$ as the mean-square closure of scores for the nuisance parameter η along all regular parametric submodels. The efficient score s_{ϕ}^{eff} for the parameter ϕ in

the model $\mathcal{M}$ is the orthogonal projection of the score s_ϕ for ϕ onto the ortho-complement $\Lambda^\perp$ to the nuisance tangent space Λ in the Hilbert space $\mathcal{L}_2 \equiv \mathcal{L}_2(\mathcal{F}_0)$ of mean zero functions with inner product $\mathrm{E}_{F_0}(T_1^T T_2)$, where F_0 is the distribution function that generated the data (Bickel et al., 1998).

Define the restricted mean model as $\mathcal{M}_{RM} = \{F_0 : \mathrm{E}[Y|\mathbf{X}] = \exp(\alpha_0 \mathbf{1}_k + \mathbf{X}\beta_0)\}$, $\theta_0 = (\alpha_0, \beta_0)$ and let $D_\beta(\mathbf{X}) = \frac{\partial \mu(\mathbf{X};\theta_0)}{\partial \beta^T}$. Bickel et al. (1998) gives the set of all influence functions for β_0 in the restricted mean model $\mathcal{M}_{RM}$ is given by:

$$\Lambda_{RM}^\perp = \left\{\varphi(\mathbf{X}) = \mathrm{E}\left[A(\mathbf{X})D_\beta(\mathbf{X})\right]^{-1} A(\mathbf{X})\epsilon : A(\mathbf{X}) \text{ arbitrary}\right\}$$

As stated before, we treat the baseline risk as a nuisance parameter in our semiparametric model. Therefore, the nuisance tangent space Λ_{RM} needs to additionally span the space of scores for α_0. In other words, $\Lambda = \Lambda_{RM} + \Lambda_\alpha$, where Λ_α is the closed linear space spanned by scores for α_0 along all regular parameteric submodels, or $\Lambda^\perp = \Lambda_{RM}^\perp \cap \Lambda_\alpha^\perp$, where Λ is the nuisance tangent space of the semiparametric model in which the baseline risk is a nuisance parameter. Using this result, one can characterize the set of influence functions for any regular and asymptotically linear estimator of β_0 in the semiparametric model that treats α_0 as a nuisance parameter. Proofs of all the following results are provided in Section A.1.1.

Result 1: The set of all influence functions of β_0 can be characterized by the set:

$$\Lambda^\perp = \left\{ \varphi(\mathbf{X}) = E\left[A(\mathbf{X})D_\beta(\mathbf{X})\right]^{-1} A(\mathbf{X})\epsilon : \begin{array}{l} A(\mathbf{X}) = h(\mathbf{X}) - \frac{E[h(\mathbf{X})\mu(\mathbf{X};\theta_0)]}{E[\mu^T(\mathbf{X};\theta_0)\mu(\mathbf{X};\theta_0)]}\mu^T(\mathbf{X};\theta_0), \\ h(\mathbf{X}) \text{ arbitrary} \end{array} \right\}$$

This implies that for any choice of $h(\mathbf{X})$, $U(h;\mathbf{X}) = A(\mathbf{X})\epsilon$ can be used as an estimating equation and the resulting estimator has influence function belonging to $\Lambda^\perp$.

Given that we have characterized the set of all influence functions, a result due to Bickel et al. (1998) states that, under certain regularity conditions, any regular and asymptoti-

cally linear estimator of β_0 that can be obtained by solving an estimating equation has an influence function belonging to $\Lambda^{\perp}$ and asymptotic distribution given by:

$$\sqrt{n}(\widehat{\beta} - \beta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varphi(\mathbf{X}) + o_p(1)$$

Standard application of the central limit theorem implies:

$$\sqrt{n}(\widehat{\beta} - \beta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \mathrm{E}[\varphi^{\otimes 2}]) \tag{2.1}$$

As we now show, the benefit of treating the log-baseline risk as a nuisance parameter in a semiparametric model is that solving an estimating equation for β_0 whose influence function belongs to $\Lambda^{\perp}$ is robust to misspecification of the baseline risk $exp(\alpha_0)$.

Result 2: Consider any $U(h; \mathbf{X}, \alpha_0, \beta_0)$ as defined in Result 1, and replace the log-baseline risk α_0 with any arbitrary value α. Then,

$$\mathrm{E}\left[U(h; \mathbf{X}, \alpha, \beta_0)\right] = 0$$

Result 2 implies that we have a set of unbiased estimating equations for β_0 that are robust to misspecification of α_0; therefore, a working estimate of the baseline risk can be used in place of the true baseline risk, and the resulting estimators are regular and asymptotically linear with influence functions belonging to $\Lambda^{\perp}$. The estimator provided for independent outcomes in Section 2.2.1 has influence function belonging to $\Lambda^{\perp}$ by taking $h(\mathbf{X}) = D_{\beta}^{T}(\mathbf{X}) V_{ind}^{-1}(\mathbf{X}) - \frac{\mathrm{E}\left[D_{\beta}^{T}(\mathbf{X}) V_{ind}^{-1}(\mathbf{X}) \mu(\mathbf{X}|\theta_0)\right]}{\mathrm{E}\left[\mu^{T}(\mathbf{X}|\theta_0) V_{ind}^{-1}(\mathbf{X}) \mu(\mathbf{X}|\theta_0)\right]} \mu^{T}(\mathbf{X}|\theta_0) V_{ind}^{-1}(\mathbf{X})$, where $V_{ind}(\mathbf{X}) = diag\{\mu(\mathbf{X}|\theta_0)(1 - \mu(\mathbf{X}|\theta_0))\}$ and remains robust to misspecification of the baseline risk for clustered outcomes. However, the estimator provided for independent outcomes is inefficient in the setting of clustered outcomes because it fails to consider the covariance structure between the clustered outcomes.

Result 3: The efficient score for β_0 in $\mathcal{M}$ is given by $U(h^{eff};\mathbf{X})$ with

$$h^{eff} = D_\beta^T(\mathbf{X})V^{-1}(\mathbf{X}) - \frac{E\left[D_\beta^T(\mathbf{X})V^{-1}(\mathbf{X})\mu(\mathbf{X}|\theta_0)\right]}{E\left[\mu^T(\mathbf{X}|\theta_0)V^{-1}(\mathbf{X})\mu(\mathbf{X}|\theta_0)\right]}\mu^T(\mathbf{X}|\theta_0)V^{-1}(\mathbf{X})$$

where $V(\mathbf{X}) = E[\epsilon\epsilon^T|\mathbf{X}]$.

The efficient score $U(h^{eff};\mathbf{X})$ given in Result 3 can be used as an estimating equation. The resulting estimator $\widehat{\beta}^{eff}$ is efficient in large samples and has asymptotic distribution given by Equation 2.1. In practice, estimation of the nuisance parameters (α_0 and $V^{-1}(\mathbf{X})$) is needed. We have already shown in Result 2 that any estimating equation for β_0 whose influence function belongs to $\Lambda^\perp$ is robust to misspecification of the log-baseline risk; as a direct result, the efficient score $U(h^{eff};\mathbf{X})$ is robust to misspecification of the log-baseline risk. Further, estimating equations for β_0 given by $\Lambda^\perp$ do not depend on the covariance structure $V(\mathbf{X})$ for unbiasedness. Therefore, any estimate of $V(\mathbf{X})$ can be used in $U(h^{eff};\mathbf{X})$ and the resulting estimator still has influence function belonging to $\Lambda^\perp$.

To construct the efficient estimate of the log risk ratio β_0, we will use the efficient score in an estimating equation. Specifically, let $\widehat{\beta}^{eff}$ be the solution to:

$$\sum_{i=1}^{n} U(h^{eff};\mathbf{X}_i, Y_i) = 0 \tag{2.2}$$

A theorem due to Bickel et al. (1998) states that for any initial $n^{1/2}$-consistent estimator of β_0, an efficient estimator can be constructed by a one-step update in the direction of the estimated efficient score using:

$$\widehat{\beta}^{eff} = \widehat{\beta} - \left[\sum_i \dot{\widehat{s}}_\beta^{eff}\right]^{-1} \sum_i \widehat{s}_\beta^{eff}$$

where $\widehat{s}_\beta^{eff}$ is an empirical version of s_β^{eff} (and $\sum_i \dot{\widehat{s}}_\beta^{eff}$ is an empirical estimator of the expected derivative of the efficient score) obtained by replacing all expectations by their

empirical counterpart, with β_0 estimated by $\widehat{\beta}$ and $\exp(\alpha_0)$ estimated by the plug-in estimator $\sum_i \mathbf{1}_k^T \mathbf{Y}_i \exp(-\mathbf{X}_i\widehat{\beta})$. Bickel et al. (1998) also states under standard regularity conditions, $n^{1/2}(\widehat{\beta}^{eff} - \beta_0)$ is asymptotically normal with mean zero and variance given as before.

In practice, each expectation is replaced with its empirical counterpart, so that $\widehat{\beta}^{eff}$ is simple to calculate. One can use the estimate provided for independent outcomes as an initial $\widehat{\beta}$; however, based on our simulations in Section 3.3, a better choice is to use the modified Poisson estimator. Note that the efficient estimator $\widehat{\beta}^{eff}$ is only feasible if $V(\mathbf{X})$ is known. Since this covariance function is unknown, it must be modeled.

A major contribution of this method is that it allows a researcher to capture the correlation among the clustered outcomes by modeling of $V^{-1}(\mathbf{X})$, which in turn may be used to increase the efficiency if correctly specified. Modeling the covariance structure for binary outcomes can be a challenging task. Consider the parameterization in terms of correlations proposed by Bahadur (1961). If we let $R_j = \frac{Y_j - \mu_j}{\{\mu_j(1-\mu_j)\}^{1/2}}$, $\rho_{jk} = \mathrm{corr}(Y_j Y_k) = \mathrm{E}(R_j R_k)$, $\rho_{jkl} = \mathrm{E}(R_j R_k R_l)$ and so on. Then,

$$Pr(\mathbf{Y} = \mathbf{y}) = \prod_{j=1}^{k} \mu_j^{y_j}(1-\mu_j)^{(1-y_j)} \left(1 + \sum_{j<k} \rho_{ik} r_j r_k + \sum_{j<k<l} \rho_{ikl} r_j r_k r_l + ... + \rho_{1...k} r_1 r_2 \cdots r_k\right)$$

We proceed under the common assumption that all 3^{rd} order or higher correlations are zero, so that all that must be specified to estimate $V^{-1}(\mathbf{X})$ is a working correlation structure, $\mathbf{R}(\boldsymbol{\rho})$. Since the model does not put any restriction on $V^{-1}(\mathbf{X})$, we additionally allow for a dispersion parameter ϕ, and $\widehat{V}(\mathbf{X}_i) = \phi \mathbf{A}_i^{1/2} \mathbf{R}(\boldsymbol{\rho}) \mathbf{A}_i^{1/2}$, where $\mathbf{A}_i = \mathrm{diag}[\widehat{\mu}_{\mathrm{i}}(1 - \widehat{\mu}_{\mathrm{i}})]$. Common choices of correlation structures include exchangeable, autoregressive, and unstructured and details of the choices and estimation of correlation parameters can be found in Liang and Zeger (1986). As a note, in theory $\phi = 1$, but we have found that allowing it be estimated from the data improves finite sample variance estimation.

2.3 Additional results and simulation

2.3.1 An alternate efficient estimator

Estimation of $\widehat{\beta}^{eff}$ depends on $\widehat{\mathbf{A}}_{ij}^{1/2} = [\widehat{\mu}_{ij}(1-\widehat{\mu}_{ij})]^{1/2}$ through the covariance function, which is only defined for $0 \leq \widehat{\mu}_{ij} \leq 1$. As such, the efficient estimator may run into convergence issues if the estimated risks are not bounded by 1. To get around such a problem, we adopt the method proposed by Tchetgen Tchetgen (2012). Specifically, let

$$logit(\mu_{ij}) = logit(\exp\left(\alpha + \mathbf{X}_{i(j)}\beta_0\right))$$

Then, ignoring knowledge about the functional form of the predicted risk, fit the model:

$$logit(\mu_{ij}) = \xi(\mathbf{X}_{i(j)}\beta_0)$$

where $\xi(\cdot)$ is an unrestricted function, and $\mathbf{X}_{i(j)}\beta_0$ is replaced with the initial estimate $\mathbf{X}_{i(j)}\widehat{\beta}$. Any nonparametric technique can be used to approximate $\xi(\cdot)$ including polynomial series, kernel smoothing, wavelet regression, or spline regression (Wasserman, 2005; Friedman et al., 2008). Let $\widehat{\xi_{ij}} = \widehat{\xi}(\mathbf{X}_{i(j)}\widehat{\beta})$ denote such an estimator, and the resulting $\widetilde{\mu}_{ij} = expit\left\{\widehat{\xi}_{ij}\right\}$ is used in the place of μ_{ij} in the updating of $\widehat{\beta}^{eff}$.

Here, we briefly illustrate that polynomial series regression does not change the efficiency of the resulting estimator. Let $\phi_k(M_i) = M_i^k$ for $k = 1, ..., K$. Then, for fixed K, let $\widetilde{p}_i$ denote the predicted probabilities obtained by standard logistic regression of Y_i on $\{\phi_k(M_i) : k \leq K\}$ using the data $\{(M_i, Y_i) : i = 1, ..., n\}$. A result due to Hirano et al. (2003) implies that since $\xi(\cdot)$ has at least four bounded derivatives, setting $K = Cn^{1/6}$ for some constant C is sufficient for the resulting estimator $\widetilde{\mu}_i$ to converge to μ_i at rates no slower than $n^{1/4}$, and the resulting estimator $\widetilde{\beta}^{eff}$ of β_0 is semiparametric efficient.

2.3.2 A more general model

All previous results were derived for the model that assumes a common baseline risk for observations within a cluster, but easily extend to a model that allows for different baseline risks. Such models are useful in the context of repeated measures over time (i.e. longitudinal data), and allow for the model to capture the risk changing over time.

As before, let $\mathbf{Y}_i$ be a k-dimensional response vector and $\mathbf{X}_i$ be a (kxq) matrix of covariates for $i = 1, \ldots, n$. Consider the semiparametric model where the only restriction is

$$\mathrm{E}\left[\mathbf{Y}|\mathbf{X}\right] = \mu(\mathbf{X}|\alpha_0, \beta_0) = \exp\left(\alpha_0 + \mathbf{X}\beta_0\right)$$

where β_0 is a q-dimensional parameter of interest and α_0 is a k-dimensional vector of log-baseline risks. Following the same development as before, it can be shown that the set of influence functions for β_0 treating the vector of baseline risks α_0 as a nuisance parameter are of the form:

$$\Lambda^{\perp} = \left\{ \begin{array}{ll} \varphi(\mathbf{X}) & = \mathrm{E}\left[A(\mathbf{X})D_{\beta}(\mathbf{X})\right]^{-1} A(\mathbf{X})\epsilon, \\ & A(\mathbf{X}) = h(\mathbf{X}) - \mathrm{E}\left[h(\mathbf{X})M(\mathbf{X};\theta_0)\right] \mathrm{E}\left[M^T(\mathbf{X};\theta_0)M(\mathbf{X};\theta_0)\right]^{-1} M^T(\mathbf{X};\theta_0), \\ & h(\mathbf{X}) \text{ arbitrary} \end{array} \right\}$$

where $D_{\beta}(\mathbf{X}) = \frac{\partial \mu(\mathbf{X};\theta_0)}{\partial \beta^T}$ and $M(\mathbf{X};\theta_0) = diag(\mu(\mathbf{X};\theta_0))$.

This set contains influence functions of all regular and asymptotically linear estimators of β_0 when the baseline risk is arbitrarily flexible. As such, this set is contained in the set of influence functions derived in Result 1 because assuming a common baseline risk is a more restrictive model. Similarly (but not exclusively), this set could also be used to construct regular and asymptotically linear estimators of β_0 in the context of longitudinal data where the baseline risk is indexed by time, $\alpha(t)$.

2.3.3 Simulations

In this section, we empirically verify the efficiency of the proposed estimator, and its robustness to misspecification of the covariance structure. We compare three estimators: (1) the estimator of Tchetgen Tchetgen (2012) which ignores possible correlation of the clustered outcomes; (2) the modified Poisson approach assuming an exchangeable correlation structure; and (3) our proposed estimator $\widehat{\beta}^{eff}$ assuming an exchangeable correlation structure.

The data is generated in a manner to reflect a cluster randomized trial for a binary treatment, and is generated as follows: (1) for each independent cluster i, generate $\mathbf{X}_i$ as $q-1$ normal random vectors and a vector of treatment indicator variables; and (2) generate the $k-$dimensional response $\mathbf{Y}_i$ such that $log(\mathrm{E}[Y_i|\mathbf{X}]) = \alpha_0 + \mathbf{X}_i\beta_0$ with correlation structure given by $\mathbf{R}$. The baseline risk was chosen to be 0.37. Various relative risks and two correlation structures were considered. First, the exchangeable correlation structure assumes all pairwise correlations between observations within a cluster are equal to ρ. This structure is widely used in practice and is useful in capturing the overall correlation within a cluster. The second correlation structure we consider mimics what might be expected if the clusters are households where the first two observations in each cluster are the parents and the remaining observations are the children. This household correlation structure is given by:

$$\begin{pmatrix} 1 & 0.05 & 0.1 & 0.1 & 0.1 \\ 0.05 & 1 & 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 1 & 0.3 & 0.3 \\ 0.1 & 0.1 & 0.3 & 1 & 0.3 \\ 0.1 & 0.1 & 0.3 & 0.3 & 1 \end{pmatrix} \tag{2.3}$$

Table 2.1 provides the absolute bias and mean squared error of each estimator for estimating the relative risk of the binary treatment when there are 1000 clusters of size 5 and the true correlation structure is either exchangeable with $\rho = 0.3$ or the household struc-

ture given in Equation 2.3. Recall that the working correlation structure for the modified Poisson and the efficient estimator is assumed to be exchangeable. The estimator that assumes independent observations has the highest mean squared error under each value of the relative risk, and the efficient estimator has the smallest mean squared error. These results are as anticipated; accounting for the correlation in the outcome improves the efficiency of both the modified Poisson and the efficient estimator. Although the modified Poisson approach accounts for correlation, it is inefficient due to misspecification of the covariance structure (due to the misspecification of the distribution). The efficient estimator correctly models this covariance structure, and as a result has the smallest mean squared error.

Consider the results when the relative risk of the binary treatment is 1.05 in Table 2.1 under the exchangeable correlation structure; we note that the three estimators have approximately the same absolute bias (2.98x10^{-3}, 2.67x10^{-3}, and 2.89x10^{-3}), but that the efficient estimator has the smallest mean squared error of 1.93x10^{-3} compared to 2.61x10^{-3} and 2.00x10^{-3}. Moving to the case where the relative risk of the binary treatment is 2, accounting for the correlation in the outcome dramatically reduces the bias, with the bias of the estimator that assumes independence equal to 6.18x10^{-3} and that of the efficient estimator equal to 0.12x10^{-3}.

Consider the situations in Table 2.1 where the true correlation structure is the household structure given in Equation 2.3. Here, the modified Poisson and efficient estimator incorrectly assume that the working correlation structure is exchangeable, but still show a reduction in mean squared error when compared to the estimator that assumes independence. The same patterns are observed under the misspecification of the covariance structure as were observed under the correct specification, with the estimator that assumes independent observations having the highest mean squared error under each value of the relative risk. In each case, the efficient estimator has smaller mean squared error than the estimator that assumes independent observations. Further, the bias of the efficient estimator remains small under the misspecification of the correlation structure. Under

Table 2.1: Results of simulation study when estimating the relative risk of a binary covariate

		Independent		Modified Poisson		Efficient		
True CS	**Relative Risk**	**MSE**	**Bias**	**MSE**	**Bias**	**MSE**	**Bias**	**Coverage**
Exch.	1	2.81	1.52	2.14	2.08	2.09	1.82	94.5
	1.05	2.61	2.98	2.00	2.67	1.93	2.89	95.1
	1.5	3.60	2.58	2.88	0.43	2.83	0.37	94.7
	2	4.57	6.18	3.74	0.41	3.67	0.12	96.2
Household	1	2.05	0.75	1.99	1.18	1.91	0.93	94.5
	1.05	2.16	3.68	2.09	2.56	1.96	2.57	95.6
	1.5	2.77	5.46	2.68	0.07	2.53	1.27	95.8
	2	3.89	10.48	3.53	3.29	3.58	1.35	95.0

Bias (10^{-3}) and mean square error (10^{-3}) of the modified Poisson approach and the efficient approach for estimating the relative risk of a binary covariate when there are 1000 clusters of size 5 under an exchangeable working correlation structure. The true correlation structure is either exchangeable with $\rho = 0.3$ or the household structure given in Equation 2.3.

the case when the relative risk of the binary treatment is 2, the efficient estimator has a bias and mean squared error of $1.35\text{x}10^{-3}$ and $3.58\text{x}10^{-3}$, respectively, while the estimator assuming independence has a larger bias and mean squared error at $10.48\text{x}10^{-3}$ and $3.89\text{x}10^{-3}$, respectively.

Table 2.2 is a reproduction of Table 2.1 but for a continuous covariate in place of the binary treatment. The results follow a similar pattern.

The results of these simulations verify that the proposed efficient estimator reduces mean squared error of the estimated risk ratios across a variety of simulated scenarios. All estimators considered in this simulation study are consistent and provide asymptotically valid inference. However, it appears that accounting for clustering in the outcomes reduces finite sample bias.

2.4 Application: Young Citizens Data

We applied our proposed estimator for the risk ratio to data from the *Young Citizens* study (Kamo et al., 2008). The trial involved a behavioral intervention designed to train children aged 10-14 years to educate their communities about HIV. The study involved 30 communities that were paired based on a clustering algorithm incorporating demographics, and one community in each pair randomly assigned treatment group with the other assigned to the control group. Residents within each community were surveyed post-intervention to determine their beliefs about the ability to children to teach the community about HIV. The primary outcome of this study was a composite scored reflecting the strength of this belief. However, to illustrate our estimator, we chose to consider a secondary outcome of the study, specifically the residents' beliefs regarding whether or not the AIDS problem was getting worse in their community (Stephens et al., 2012). This outcome was derived by collapsing a 4-point scale with values "strongly agree", "agree", "disagree", or "strongly disagree" into two values, "agree" or "disagree".

Table 2.2: Results of simulation study when estimating the relative risk of a continuous covariate

		Independent		Modified Poisson		Efficient		
True CS	**Relative Risk**	**MSE**	**Bias**	**MSE**	**Bias**	**MSE**	**Bias**	**Coverage**
Exch.	1	0.31	0.06	0.25	0.08	0.23	0.23	94.7
	1.05	0.33	1.49	0.26	0.20	0.24	0.02	94.5
	1.5	0.74	6.45	0.55	0.68	0.50	1.27	94.5
	2	1.66	12.47	1.23	0.069	1.05	1.68	95.4
Household	1	0.286	0.01	0.284	0.09	0.275	0.15	93.9
	1.05	0.27	1.76	0.28	0.87	0.24	1.13	94.8
	1.5	0.66	13.56	0.44	0.095	0.43	0.50	94.1
	2	2.01	25.66	0.818	1.69	0.816	0.57	93.0

Bias (10^{-3}) and mean square error (10^{-3}) of the modified Poisson approach and the efficient approach for estimating the relative risk of a continuous covariate when there are 1000 clusters of size 5 under an exchangeable working correlation structure. The true correlation structure is either exchangeable with $\rho = 0.3$ or the household structure given in Equation 2.3.

We estimated the risk ratio of the intervention using the efficient estimator given in Section 2.2.2 assuming an exchangeable correlation structure, the modified Poisson approach assuming an exchangeable correlation structure, and the estimator that assumes independence given in Section 2.2.1. Additionally, we estimate the odds ratio of the intervention using a GEE with a logit link and assuming an exchangeable correlation structure. In all of the estimators, we control for the baseline covariates residential or urban community, religion, ethnic group, and indicators of wealth by including the covariates into the linear predictor of the mean.

Table 2.3 provides the estimated risk ratio of the intervention, the standard error, and the 95% confidence interval for each of the estimators considered. We would like to note that standard GEE for the log-binomial model with correlated data failed to converge, and as such, a different approach must be taken to estimate the risk ratios. The outcome is not rare (∼82% responded "agree"); therefore, using odds ratios to estimate the risk ratio is not valid.

Table 2.3: Results of analysis of *Young Citizens* study

Estimator	**log(Risk ratio)**	**Std. Error**	**95% Confidence Interval**
$\widehat{\beta}^{eff}$	-0.0188	0.0375	(-0.0922 , 0.0547)
$\widehat{\beta}^{MP}$	-0.0206	0.0406	(-0.1002, 0.0590)
$\widehat{\beta}^{OR}$	-0.1222	0.2529	(-0.6179 , 0.3736)

Estimated log-risk ratio (or log-odds ratio) of the intervention, the standard error, and corresponding 95% confidence interval. $\widehat{\beta}^{eff}$ is the efficient estimator provided in Section 2.2.2 assuming an exchangeable correlation structure, $\widehat{\beta}^{MP}$ is the modified Poisson estimator assuming an exchangeable correlation structure, and $\widehat{\beta}^{OR}$ is the log-odds ratio estimated using the GEE with a logit link and assuming an exchangeable correlation structure.

The efficient estimator and that of the modified Poisson approach provide similar estimates of the log-risk ratio, -0.0188 and -0.0206, respectively, with the efficient estimator slightly smaller in magnitude. The standard error of the efficient estimator is 0.0375, compared to 0.0406 for the modified Poisson approach. This corresponds to an empirical asymptotic relative efficiency of 0.85 for the modified Poisson compared to the efficient estimator, and is reflected in by a narrowing of the confidence intervals. Neither approach

leads to significant effects at the $\alpha = 0.05$, but the results do illustrate the efficient estimator has tighter confidence intervals than that of the modified Poisson approach. Also provided in Table 2.3 is the log-odds ratio estimated using a GEE with a logit link and assuming an exchangeable correlation structure. The estimated log-odds ratio is -0.1222, illustrating that the odds ratio is not a good approximation of the risk ratio in the trial and likely overestimates the relative risk of the intervention.

2.5 Discussion

In this paper, we have proposed an efficient estimator of the risk ratio that accounts for clustering among binary outcomes. We prove that this estimator is robust to misspecification of the baseline risk, in the sense that the estimator does not directly rely on an estimate of the baseline risk for consistency, and showed that it has the smallest asymptotic variance of any regular and asymptotically linear estimator. Further, a modification of the estimator is provided that guarantees the predicted probability is bounded by 1 (a model restriction), and as a result, guarantees stable performance of the estimator.

Simulations confirm that the proposed estimator has smaller variance than estimators that assume independence and the modified Poisson approach both under correct and incorrect specification of the correlation structure. Additionally, the simulations suggest that the proposed estimator may have smaller finite sample bias in the estimation of the risk ratios when compared to estimators that assume independence. Therefore, it is important to account for correlation among clustered outcomes both to improve efficiency and to remove finite sample bias.

The gains in efficiency of the proposed estimator when compared to the modified Poisson approach are due to allowing for correct specification of the underlying data distribution. A priori, the modified Poisson approach incorrectly models the data as a Poisson distribution, leading to a misspecification of the covariance structure and ruling out the

possibility of an efficient estimator. The estimator proposed in this paper allows for correct distributional assumptions, and avoids the common drawbacks of this assumption by being robust to misspecification of the baseline risk.

3. Model averaged double robust estimation

[1]Matthew Cefalu, [1]Francesca Dominici, and [1,2]Giovanni Parmigiani

[1]Department of Biostatistics, Harvard School of Public Health
[2]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute

Abstract

We propose a new class of estimators for the average causal effect, the model averaged double robust (MA-DR) estimators, that account for model uncertainty in both the propensity score and outcome model through the use of model averaging. The MA-DR estimator is defined as a weighted average of double robust estimators, where each double robust estimator corresponds to a specific choice for the outcome model and the propensity score, respectively. The MA-DR estimators extend the desirable double robustness property by achieving consistency under the much weaker assumption that either the true propensity score model or the true outcome model be within a specified, possibly large, class of models. We provide asymptotic results and conduct a large scale simulation study that indicates the MA-DR estimator has better finite sample behavior than the usual double robust estimator. We show that the MA-DR that a priori links the propensity score and the outcome model can have 90% less variance than a double robust estimator constructed via model selection for the propensity score and the outcome model separately. Importantly, our simulation suggests that our MA-DR estimator dramatically reduces mean squared error by the largest percentage in the realistic situation where the outcome model is misspecified.

3.1 Introduction

Methods for causal inference are predicated on knowledge of the covariates necessary to satisfy the no unmeasured confounding assumption, but the exact set of covariates needed to control confounding is rarely known. Practical tools that acknowledge uncertainty in confounder selection and are robust to model misspecification are imperative for correct estimation of the average causal effect (Vansteelandt et al., 2010; Wang et al., 2012a).

Although the literature on causal inference is vast, existing methods do not account for the

uncertainty in selection of confounders, C, or in the form of the model for the treatment, X (Vansteelandt et al., 2010). For example, a wealth of methods that rely on specification of a propensity score model, $\mathrm{P}(X|C)$, for treatment assignment (e.g propensity score matching or inverse probability weighting estimators; see (Lunceford and Davidian, 2004) for a review) typically assume that both the covariates to include and the functional form of the propensity score model are known a priori.

In addition to specification of $\mathrm{P}(X|C)$, a broad class of methods for causal inference rely on the additional specification of a model for potential outcomes $\mathrm{P}(Y(x)|C)$ (Rosenbaum and Rubin, 1983, 1984; Drake, 1993), where $Y(1)$ and $Y(0)$ are the potential outcomes under each treatment. Included in this class are methods for inverse probability treatment weighted estimators that are often promoted for properties such as consistency and double robustness (Scharfstein et al., 1999; Bang and Robins, 2005; Tan, 2010). Within the class of double robust estimators, covariate and model selection is specified a priori for both the propensity score and the outcome models separately, presenting further challenges to acknowledging uncertainty with respect to the selection of the confounders and providing robustness to model misspecification. There are few tools or guidelines for model selection in double robust estimators, and many researchers take an ad-hoc approach.

One possible tool to formally account for model uncertainty in the adjustment for confounding is Bayesian model averaging (Raftery et al., 1997; Draper, 1995). These methods are based on treating the indicators of whether each confounder is included in the model as a nuisance parameter, and it has been suggested that an effect estimate can be formed by weighting the model-specific estimates (Hoeting et al., 1999), where the weights are determined by the models' posterior probabilities.

In the context of a regression model, where the goal is the estimation of the effect of X on Y adjusting for measured confounders, the use of Bayesian model averaging with non-informative priors on the models has received some criticism (Crainiceanu et al., 2008; Vansteelandt et al., 2010; Wang et al., 2012a): variable selection based on the outcome

model only prioritizes the Cs strongly associated with Y, and variable selection based on the propensity score model only prioritize the Cs that are strongly associated with X. Both these approaches can result in inefficient and biased inferences because they fail to identify the full set of necessary confounders (Brookhart, 2006; Schneeweiss et al., 2009).

Wang et al. (Wang et al., 2012a) propose a solution to this important problem for a continuous exposure X and with confounding adjustment made by introducing Cs into the regression model as covariates. Two regression equations are specified along with two vectors of inclusion indicators: (1) a linear regression model for Y given X and C (the outcome model); and (2) a linear regression model for X given C (the exposure model). They assume a priori that if a covariate C is highly predictive of the exposure X, then the same covariate C will have a large probability of being included into the outcome model. It is shown that the model averaged estimator of the effect of X on Y, obtained with this form of prior dependence between the outcome model and the exposure model has lower mean squared error than the model averaged estimator that assumes a priori that the two vectors of inclusions indicators are independent.

Accounting for model uncertainty in the context of causal inference is a widely unexplored topic. In this paper, we propose a new class of methods for estimating the average causal effect, which we call *the model averaged double robust estimators*, that formally account for model uncertainty through the use of model averaging while maintaining the desirable properties of consistency and double robustness. These methods provide valid estimation of the average causal effect that: 1) are robust to the misspecification of the model for the treatment assignment; 2) are robust to the misspecification of the model for outcome; and 3) account for the uncertainty in the selection of the confounders in both the propensity score model and in the outcome model. Importantly, we show that a model averaged double robust estimator that assumes dependence between the propensity score and the outcome model a priori and separates estimation of the model weights into two stages can reduce the mean squared error of the double robust estimator by more than 90% when compared to traditional model selection procedures.

3.2 Methods

3.2.1 A double robust estimator

Consider continuous potential outcomes $(Y(0), Y(1))$, binary treatment X, and a p-dimensional set of potential confounders C. Assume there is no unmeasured confounding (Robins et al., 2000) (also referred to as strong ignorable treatment assignment(Rosenbaum and Rubin, 1983)), so that $(Y(0), Y(1)) \perp\!\!\!\perp X|C$. Let (Y_h, X_h, C_h) be independent observations for $h = 1, ..., n$. We are interested in estimating the average causal effect:

$$\Delta = E\left[Y(1) - Y(0)\right] = E\left\{E(Y|X = 1, C) - E(Y|X = 0, C)\right\} \tag{3.1}$$

Given a model for the propensity score, $\mathrm{P}(X = 1|C) = e(C)$, and a model for the outcome under each treatment, $E(Y|X = 1, C) = m_1(C)$ and $E(Y|X = 0, C) = m_0(C)$, we define the well known parametric ($\widehat{\Delta}^p$), inverse probability weighted ($\widehat{\Delta}^{IPW}$), and double robust ($\widehat{\Delta}^{DR}$) estimators as:

$$\begin{aligned}
\widehat{\Delta}^p &= \frac{1}{n}\sum_{h=1}^{n}\{\widehat{m}_{1h} - \widehat{m}_{0h}\} \\
\widehat{\Delta}^{IPW} &= \frac{1}{n}\sum_{h=1}^{n}\left\{\frac{Y_h X_h}{\widehat{e}_h} - \frac{Y_h(1 - X_h)}{1 - \widehat{e}_h}\right\} = \frac{1}{n}\sum_{i=1}^{n}\frac{X_h - \widehat{e}_h}{\widehat{e}_h(1 - \widehat{e}_h)}Y_h \\
\widehat{\Delta}^{DR} &= \frac{1}{n}\sum_{h=1}^{n}\left\{\frac{Y_h X_h - (X_h - \widehat{e}_h)\widehat{m}_{1h}}{\widehat{e}_h} - \frac{Y_h(1 - X_h) + (X_h - \widehat{e}_h)\widehat{m}_{0h}}{1 - \widehat{e}_h}\right\}
\end{aligned} \tag{3.2}$$

where $\widehat{m}_{1h}$, $\widehat{m}_{0h}$, and $\widehat{e}_h$ are the estimated outcomes and propensity score for individual h under model $m_1(C)$, $m_2(C)$, and $e(C)$, respectively. To simplify the model averaging arguments in the next section, note that $\widehat{\Delta}^{DR}$ can be decomposed into $\widehat{\Delta}^{IPW}$, $\widehat{\Delta}^p$, and a third estimator $\widehat{\Delta}^{PIPW}$.

$$\begin{aligned}\widehat{\Delta}^{DR} &= \frac{1}{n}\sum_{h=1}^{n}\left[\widehat{m}_{1h} - \widehat{m}_{0h} + \frac{X_h - \widehat{e}_h}{\widehat{e}_h(1-\widehat{e}_h)}Y_h - \frac{X_h - \widehat{e}_h}{\widehat{e}_h(1-\widehat{e}_h)}\widehat{m}_{X_h}\right] \\ &= \widehat{\Delta}^{p} + \widehat{\Delta}^{IPW} - \widehat{\Delta}^{PIPW}\end{aligned}$$

where $\widehat{\Delta}^{PIPW}$ is a parametric inverse probability weighted estimator and $\widehat{m}_{X_h} = \widehat{m}_{1h}$ if $X_h = 1$ and $\widehat{m}_{X_h} = \widehat{m}_{0h}$ otherwise. Observe that $\widehat{\Delta}^{p}$ only depends on the outcome model, $\widehat{\Delta}^{IPW}$ only depends on the propensity score model, and $\widehat{\Delta}^{PIPW}$ depends on both.

The model for the propensity score and the outcome under each treatment can be selected in any number of ways. A researcher may rely on expert knowledge to decide both the functional form and the covariates to include in each model, or may rely on a model selection procedure that chooses the best model from a set of candidate models. For the remainder of this paper, we will refer to $\widehat{\Delta}^{MS}_{DR}$ as the "model selected double robust estimator" in which both the propensity score and the outcome model have been selected independently using BIC (Schwarz, 1978) as a model selection procedure.

3.2.2 Model averaged double robust estimator

Let $\mathcal{M}^{ps} = \left\{\mathcal{M}^{ps}_1, \mathcal{M}^{ps}_2, ..., \mathcal{M}^{ps}_{M_{ps}}\right\}$, $\mathcal{M}^{0} = \left\{\mathcal{M}^{0}_1, \mathcal{M}^{0}_2, ..., \mathcal{M}^{0}_{M_0}\right\}$, and $\mathcal{M}^{1} = \left\{\mathcal{M}^{1}_1, \mathcal{M}^{1}_2, ..., \mathcal{M}^{1}_{M_1}\right\}$ be finite collections of models for $\mathrm{P}(X = 1|C)$, $E(Y|X = 0, C)$, and $E(Y|X = 1, C)$, respectively. For example, the collection of models for the propensity score $\mathcal{M}^{ps}$ could consist of logistic regression models with all subsets of C as linear predictors. Let $\mathcal{M}^{om} = \mathcal{M}^{1} \times \mathcal{M}^{0}$ denote all combinations of models in $\mathcal{M}^{1}$ and $\mathcal{M}^{0}$. Further, define $\widehat{\Delta}^{DR}_{ij}$ as the double robust estimator corresponding to the models $\mathcal{M}^{ps}_i$ and $\mathcal{M}^{om}_j$. We define the model average double robust estimator as:

$$\widehat{\Delta}_{DR}^{MA} = \sum_{ij} p_{ij} \widehat{\Delta}_{ij}^{DR} \tag{3.3}$$

where $p_{ij} = \mathrm{P}(\mathcal{M}_i^{ps}, \mathcal{M}_j^{om} | \mathcal{D})$ is the joint posterior probability of models $\mathcal{M}_i^{ps}$ and $\mathcal{M}_j^{om}$. We expand the estimator based on the decomposition in the previous section. Let $\widehat{\Delta}_i^{IPW}$, $\widehat{\Delta}_j^{p}$, and $\widehat{\Delta}_{ij}^{PIPW}$ be the inverse probability weighted estimator, the parametric estimator, and the parametric inverse probability weighted estimator for the corresponding models $\mathcal{M}_i^{ps}$ and $\mathcal{M}_j^{om}$. Then,

$$\widehat{\Delta}_{DR}^{MA} = \sum_{i} p_{i\bullet} \widehat{\Delta}_i^{IPW} + \sum_{j} p_{\bullet j} \widehat{\Delta}_j^{p} - \sum_{ij} p_{ij} \widehat{\Delta}_{ij}^{PIPW} \tag{3.4}$$

where $p_{i\bullet} = \sum_j p_{ij}$ and $p_{\bullet j} = \sum_i p_{ij}$. Note that Equation 3.4 has a model averaged term for the inverse probability weighted, parametric, and parametric inverse probability weighted estimators. The variance of the model averaged double robust estimator can be estimated using standard resampling methods (e.g. bootstrap; see (Efron and Tibshirani, 1993)).

3.2.3 Prior and posterior model probabilities

To complete the specification of the model averaged double robust estimator, a prior distribution on the model class must be assumed. We will return to choices of priors momentarily, but first let A_i be the prior odds of $\mathcal{M}_i$ versus some other model $\mathcal{M}_1$ that both belong to some model class $\mathcal{M}$. Then, the posterior probability of model $\mathcal{M}_i$ is given by:

$$\mathrm{P}(\mathcal{M}_i | \mathcal{D}) = \frac{A_i B_{i1}}{\sum_{j:\mathcal{M}_j \in \mathcal{M}} A_j B_{j1}} \tag{3.5}$$

where B_{i1} is the Bayes factor for model $\mathcal{M}_i$ against another model $\mathcal{M}_1$. Bayes factors and their estimates are well studied, and there is extensive literature on the subject (Smith and Spiegelhalter, 1980; Nishii, 1984; Kass and Raftery, 1995; Konishi and Kitagawa, 1996). Among the properties of Bayes factors is consistency for model selection, which is a necessary component for consistency of $\widehat{\Delta}_{DR}^{MA}$ as seen in Section 3.2.4. A well known and popular estimate of Bayes factors is based on BIC (Schwarz, 1978) and allows us to estimate posterior model probabilities with ease.

Returning to the specification of a prior distribution for the model space, the simplest choice is to assume that all models are equally likely a priori. This corresponds to assuming that the prior odds of each model is 1, and that the form of the propensity score and the outcome model are independent. Therefore, the resulting model averaged double robust estimator is given by:

$$\widehat{\Delta}_{DR}^{MA-i} = \sum_i p_i \widehat{\Delta}_i^{IPW} + \sum_j q_j \widehat{\Delta}_j^{p} - \sum_{ij} p_i q_j \widehat{\Delta}_{ij}^{PIPW} \tag{3.6}$$

where $p_i = \mathrm{P}(\mathcal{M}_i^{ps}|\mathcal{D})$ and $q_j = \mathrm{P}(\mathcal{M}_j^{om}|\mathcal{D})$. Notice that because of the prior independence, the posterior probabilities of the propensity score and outcome models are also independent. Therefore, $\mathrm{P}(\mathcal{M}_i^{ps}|\mathcal{D})$ and $\mathrm{P}(\mathcal{M}_j^{om}|\mathcal{D})$ can be computed separately using readily available software, and the model averaged double robust estimator is straightforward to calculate as given in Equation 3.6.

However, efficiency can be gained through the use of a prior on the model space that identifies confounders (Cs that are associated with both treatment and outcome) for use in the propensity score model. Under the prior independence assumption, the posterior model probability of the propensity score only prioritizes models in which the Cs are strongly associated with X and ignores all relationships with Y. The current literature in causal inference suggests that inclusion of covariates that are only related to the exposure into a propensity score model adds to the variance of the resulting double robust estimator

(Rubin et al., 1997; Brookhart, 2006).

In this light, we propose an alternative formulation of the prior distribution on the model space that links the propensity score model to the outcome model through prior model dependence. First, let the prior odds of propensity score model $\mathcal{M}_i^{ps}$ to $\mathcal{M}_1^{ps}$ conditional on the outcome model $\mathcal{M}_j^{om}$ be such that:

$$\frac{\mathrm{P}(\mathcal{M}_i^{ps}|\mathcal{M}_j^{om})}{\mathrm{P}(\mathcal{M}_1^{ps}|\mathcal{M}_j^{om})} = \begin{cases} 1, & \text{if } \mathcal{M}_i^{ps} \subset \mathcal{M}_j^{om} \\ 0, & \text{otherwise} \end{cases} \tag{3.7}$$

where $\mathcal{M}_i^{ps} \subset \mathcal{M}_j^{om}$ indicates that the systematic component of $\mathcal{M}_i^{ps}$ is a subset of the systematic component of $\mathcal{M}_j^{om}$. We refer to the 'systematic component' of a model as the specification of its linear predictor, so that there is no issue with the exposure being binary while the outcome is continuous, and we assume $\mathcal{M}^{ps}$ and $\mathcal{M}^{om}$ contain models with the same nested systematic components. We then choose a reference propensity score model $\mathcal{M}_1^{ps}$ such that $\mathcal{M}_1^{ps} \subset \mathcal{M}_j^{om}$ for all j. The reference model is either a null model or a model that includes confounders that are strictly required regardless of the inclusion of the other confounders. Further, assume that the prior distribution on the outcome model space is uniform. We will denote the model averaged double robust estimator using this prior as $\widehat{\Delta}_{DR}^{MA-d}$. The posterior model probabilities used in the construction of $\widehat{\Delta}_{DR}^{MA-d}$ can be estimated by first finding the prior odds of each model combination under the prior model dependence given by Equation 3.7 and then using Equation 3.5 to find the posterior model probabilities.

The prior model dependency given by Equation 3.7 forces the set of potential confounders included in the propensity score model to be a subset of those that are included in the outcome model. In other words, the prior probability of excluding a variable from the propensity score model given that it is excluded from the outcome model is one, and the prior probability of including a variable in the outcome model given that it is in the propensity score model is one. This type of restriction is supported by the current litera-

ture on propensity scores (Rubin et al., 1997; Brookhart, 2006), and is related to the priors for continuous exposure introduced by Wang et al. (Wang et al., 2012a).

Our motivation for this prior distribution on the model space was to identify the set of potential confounders that should be included into the propensity score model based on the fact that they are associated with both treatment and outcome, instead of being associated with treatment only. In other words, the prior dependency given in Equation 3.7 gives zero weight a priori to propensity score models having a systematic component that is not included in the outcome model. However, the estimation of the joint posterior model probability $\mathrm{P}(\mathcal{M}_i^{ps}, \mathcal{M}_j^{om}|\mathcal{D})$ based on this prior has the additional undesirable property that it allows feedback from the propensity score into the outcome model. This feedback is such that the posterior model probabilities will favor outcome models that include any of the potential confounders that are associated with either X or Y, and some efficiency is lost by including potential confounders that are only associated with X into the outcome model.

We will cut the feedback from propensity score into the outcome model with the goal of improving the efficiency of the model averaged double robust estimator through the use of a two-stage approach for calculating the model weights. The two-stage approach for calculating the model weights and the resulting model averaged double robust estimator proceeds as follows:

1. Estimate the marginal posterior of the outcome model, $q_j = \mathrm{P}(\mathcal{M}_j^{om}|\mathcal{D})$, assuming a uniform prior on the outcome model space and ignoring the specification of the propensity score model

2. Estimate the posterior of the propensity score model conditional on the outcome model, $\mathrm{P}(\mathcal{M}_i^{ps}|\mathcal{M}_j^{om}, \mathcal{D})$, using the prior model dependence given by Equation 3.7

3. Multiply the estimates from Stage 1 and 2 to find the joint model weight of the propensity score and outcome model, $p_{ij}^* = q_j \mathrm{P}(\mathcal{M}_i^{ps}|\mathcal{M}_j^{om}, \mathcal{D})$

4. Estimate the resulting model averaged double robust estimator $\widehat{\Delta}_{DR}^{MA-dII}$ using p_{ij}^* as the model weights. The notation II in the superscript of the estimator corresponds to the fact that we are calculating the model weights in two stages.

The model weights under this two-stage approach can be easily calculated because they are a transformation of the model probabilities assuming a uniform prior on the model space. First, the outcome model probabilities $q_j = \mathrm{P}(\mathcal{M}_j^{om}|\mathcal{D})$ in Step 1 are simply the model probabilities assuming a uniform prior on the model space. For the estimation of $\mathrm{P}(\mathcal{M}_i^{ps}|\mathcal{M}_j^{om}, \mathcal{D})$ in Step 2, note that conditional on each outcome model, the prior odds for the propensity score models are uniform for models that meet the restriction $\mathcal{M}_i^{ps} \subset \mathcal{M}_j^{om}$. Simple implementation transforming either Bayes' factors, BIC, or model probabilities under a uniform prior on the model space to model weights using the two-stage approach is available.

The difference between the two-stage model weights given by p_{ij}^* and the proper posterior model probabilities used in the estimator $\widehat{\Delta}_{DR}^{MA-d}$ is that the two-stage approach forces the marginal outcome model weights to be equal to the marginal posterior outcome model probabilities under a uniform prior on the model space. More specifically, the estimation of q_j in Stage 1 of the two-stage method does not correspond to the true marginal posterior $P(\mathcal{M}_j^{om}|\mathcal{D})$, while the estimation of $\mathrm{P}(\mathcal{M}_i^{ps}|\mathcal{M}_j^{om}, \mathcal{D})$ in Stage 2 does correspond to the true conditional posterior.

3.2.4 Asymptotic properties of $\widehat{\Delta}_{DR}^{MA}$

All of the results of this section require consistency of the posterior model probabilities. As stated in Section 3.2.3, the Bayes factor and its BIC approximations are consistent (Kass and Raftery, 1995).

We will show that if either the true propensity score model is contained in $\mathcal{M}^{ps}$ or the true outcome models are contained in $\mathcal{M}^{om}$ and the posterior model probabilities are consis-

tent for selecting the true models, then we have that $\widehat{\Delta}_{DR}^{MA}$ is consistent for the average casual effect defined in Equation 3.1. This result implies that we have added another layer of robustness to the double robust estimator, as we only need the true models to be in the collection of models. All $\rightarrow$ indicate limits as $n \rightarrow \infty$, and $\xrightarrow{p}$ indicates convergence in probability while $\xrightarrow{\mathcal{L}}$ indicates convergence in distribution.

Lemma 1. *Assume there is no unmeasured confounding, independent observations, and that $\mathcal{M}^{om}$ contains the true model, $\mathcal{M}_1^{om}$, for both $E(Y|X=1,C)$ and $E(Y|X=0,C)$. If $p_{\bullet 1} = \sum_i \mathrm{P}(\mathcal{M}_i^{ps}, \mathcal{M}_1^{om}|\mathcal{D}) \xrightarrow{p} 1$, then*

$$\widehat{\Delta}_{DR}^{MA} \xrightarrow{p} \Delta$$

Lemma 2. *Assume there is no unmeasured confounding, independent observations, and that $\mathcal{M}^{ps}$ contains the true model, $\mathcal{M}_1^{ps}$, for $\mathrm{P}(X=1|C)$. If $p_{1\bullet} = \sum_j \mathrm{P}(\mathcal{M}_1^{ps}, \mathcal{M}_j^{om}|\mathcal{D}) \xrightarrow{p} 1$, then*

$$\widehat{\Delta}_{DR}^{MA} \xrightarrow{p} \Delta$$

Theorem 1. *Assume there is no unmeasured confounding, independent observations, and let $\mathcal{M}^{om}$ and $\mathcal{M}^{ps}$ be collections of models. If,*

(1) $\mathcal{M}^{om}$ contains the true models, $\mathcal{M}_1^{om}$, for both $E(Y|X=1,C)$ and $E(Y|X=0,C)$, and $p_{\bullet 1} = \sum_i \mathrm{P}(\mathcal{M}_i^{ps}, \mathcal{M}_1^{om}|\mathcal{D}) \xrightarrow{p} 1$

<u>or</u>

(2) $\mathcal{M}^{ps}$ contains the true model, $\mathcal{M}_1^{ps}$, for $\mathrm{P}(X=1|C)$, and $p_{1\bullet} = \sum_j \mathrm{P}(\mathcal{M}_1^{ps}, \mathcal{M}_j^{om}|\mathcal{D}) \xrightarrow{p} 1$

Then,

$$\widehat{\Delta}_{DR}^{MA} \xrightarrow{p} \Delta$$

Proof. These results can be verified through the use of Slutsky's theorem and standard arguments utilizing the no unmeasured confounding assumption. □

The consistency of $\widehat{\Delta}_{DR}^{MA}$ was shown here in relation to the true propensity score and outcome model. However, the requirement that $\mathcal{M}^{om}$ and $\mathcal{M}^{ps}$ contain the truth could be replaced with the requirement that $\mathcal{M}^{om}$ and $\mathcal{M}^{ps}$ contain a model that is sufficient to control confounding. No longer would the requirement be that posterior model probability be consistent for the truth, but only that the sum of the posterior model probabilities that adequately control confounding converges in probability to 1.

Next we will show that if the collection of models $\mathcal{M}^{om}$ and $\mathcal{M}^{ps}$ contain the true models and the posterior model probabilities are $\sqrt{n}$-consistent for model selection, then $\widehat{\Delta}_{DR}^{MA}$ is asymptotically equivalent to $\widehat{\Delta}_{DR}$ when the true outcome and propensity score models are known a priori and achieves the semiparametric variance bound.

Theorem 2. *Consider $\widehat{\Delta}_{DR}^{MA}$ as described by Equation 3.4. Assume no unmeasured confounding and independent observations. Let $\mathcal{M}^{om}$ and $\mathcal{M}^{ps}$ be collections of models that contain the true models for $E(Y|X=1,C)$, $E(Y|X=0,C)$, and $Pr(X=1|C)$. Let $\mathcal{M}_1^{om}$ and $\mathcal{M}_1^{ps}$ denote the true outcome and propensity score models, and let $\widehat{\Delta}_{11}^{DR}$ be the double robust estimator using $\mathcal{M}_1^{om}$ and $\mathcal{M}_1^{ps}$. Assume the usual regularity conditions so that $n^{1/2}(\widehat{\Delta}_{11}^{DR} - \Delta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \mathcal{V})$ where $\mathcal{V}$ is the semiparametric variance bound. If $p_{11} = 1 - o_p(\frac{1}{\sqrt{n}})$, then*

$$n^{1/2}(\widehat{\Delta}_{DR}^{MA} - \Delta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \mathcal{V})$$

or

$$n^{1/2}(\widehat{\Delta}_{DR}^{MA} - \widehat{\Delta}_{11}^{DR}) = o_p(1)$$

Proof. This result can be verified by recognizing that $p_{11} = 1 - o_p(\frac{1}{\sqrt{n}})$ implies $p_{ij} = o_p(\frac{1}{\sqrt{n}})$ for $(i,j) \neq (1,1)$, and repeated application of Slutsky's Theorem. □

The restriction on the posterior model probabilities ($p_{11} = 1 - o_p(\frac{1}{\sqrt{n}})$) for this result is quite strong and is not expected to be satisfied easily. Technically speaking, all of the previous results hold for any consistent model selection procedure, whether Bayesian or frequentist. This is not surprising, as model averaging and model selection are asymptotically equivalent. However, we expect that in finite samples model averaging and model selection will differ.

3.3 Simulations

3.3.1 Set up

It is not the purpose of these simulations to verify the asymptotic properties of $\widehat{\Delta}_{DR}^{MA}$, but instead to illustrate its finite sample behavior relative to the double robust estimator using model selection for both the propensity score and the outcome model ($\widehat{\Delta}_{DR}^{MS}$). We consider the model averaged double robust estimator assuming both prior model independence and prior model dependence. Let $\widehat{\Delta}_{DR}^{MA-i}$ denote the model averaged double robust estimator that assumes prior model independence given by Equation 3.6, let $\widehat{\Delta}_{DR}^{MA-d}$ denote the model averaged double robust estimator that assumes the prior model dependence given by Equation 3.7, and let $\widehat{\Delta}_{DR}^{MA-dII}$ be the model averaged double robust estimator that assumes the prior model dependence given by Equation 3.7 and uses the two-stage approach for calculating model weights. See Table 3.1 for a description of the estimators considered in these simulations.

We use two groups of simulations. In Group 1, all effects of the potential confounders are linear in both treatment and outcome. In Group 2, we allow for interactions and non-linearities in the confounder-outcome and confounder-treatment relationships. We consider a situation where we have 5 measured potential confounders. In all of our simulations, we restrict $\mathcal{M}^{ps}$ and $\mathcal{M}^{om}$ to only include linear combinations of the potential confounders so that there are $2^5 = 32$ models for both the propensity score and the out-

Table 3.1: Description of all estimators used in the simulation study comparing double robust estimators for the average causal effect

Estimator	Description
$\widehat{\Delta}_{DR}^{MS}$	Model selected double robust estimator that chooses propensity model and the outcome model based on the BIC
$\widehat{\Delta}_{DR}^{MA-i}$	MA-DR estimator assuming prior model independence
$\widehat{\Delta}_{DR}^{MA-d}$	MA-DR estimator assuming prior model dependence defined by Equation 3.7
$\widehat{\Delta}_{DR}^{MA-dII}$	MA-DR estimator assuming prior model dependence defined by Equation 3.7 and using the two-stage approach for calculating model weights

Included is (1) the type of estimator; and (2) the choice of prior distribution for the model space. All Bayes factors are estimated using the BIC approximation.

come. Therefore, in Group 2, the true models are not contained in either $\mathcal{M}^{ps}$ or $\mathcal{M}^{om}$.

A full description of all scenarios can be found in Table 3.2 and Table 3.3. All simulations set $\beta = 1$ and use a sample size of 500 with 10,000 replications. In Group 1, we generate the data as follows: (1) $C_1, ..., C_5 \overset{iid}{\sim} N(0, 1)$; (2) $X \sim$ Bernoulli$(p = \text{expit}(C\alpha^{ps}))$; and (3) $Y \sim \text{N}(\beta X + C\alpha^{om}, 1)$. We consider different values of the unknown parameters α^{ps} and α^{om} to mimic different levels of confounding.

In Group 2, we generate the data in a similar manner, but with non-linearities in the propensity score or outcome models as follows: (1) $C_1, ..., C_5 \overset{iid}{\sim} N(0, 1)$; (2) $X \sim$ Bernoulli$(p = \text{expit}(f(C))$; and (3) $Y \sim \text{N}(\beta X + g(C), 1)$, where $f(\cdot)$ and $g(\cdot)$ are polynomial functions of C. For example, Scenario 7 is linear in the propensity score model, with $f(C) = C_1 + C_2 + C_5$, but non-linear in the outcome, with $g(C) = 0.5 \sum_{i=1}^{5} \sum_{j=1}^{5} C_i C_j$. Additional simulation scenarios and sensitivity analyses are included in Section A.2.2.

Table 3.2: Description of Group 1 in the simulation study comparing double robust estimators for the average causal effect

Scenario	$\boldsymbol{\alpha}^{\mathbf{ps}}$ **(PS model)**	$\boldsymbol{\alpha}^{\mathbf{om}}$ **(Outcome model)**
1	(0.4,0.3,0.2,0.1,0)	(0,0,0,0,0)
2	(0.5,0.5,0.1,0,0)	(0.5,0,1,0.5,0)
3	(0.1,0.1,1,1,1)	(2,2,0,0,0)
4	(0.5,0.4,0.3,0.2,0.1)	(0.5,1,1.5,2,2.5)
5	(0.5,0.4,0.3,0.2,0.1,0,0,0,0,0)	(0.5,1,1.5,2,2.5,0,0,0,0,0)

All effects of confounders are linear on both the treatment and outcome. Data is generated as follows: (1) $C_1, ..., C_5 \overset{iid}{\sim} N(0, 1)$; (2) $X \sim$ Bernoulli$(p = \text{expit}(C\alpha^{ps}))$; and (3) $Y \sim \text{N}(\beta X + C\alpha^{om}, 1)$

3.3.2 Results

Recall that the "model selected double robust estimator", $\widehat{\Delta}_{DR}^{MS}$, refers to the double robust estimator in which both the propensity score and the outcome model have been selected independently using BIC based model selection. Table 3.4 provides the percent decrease in mean squared error of the three model averaged double robust estimators defined in

Table 3.3: Description of Group 2 in the simulation study comparing double robust estimators for the average causal effect

Scenario	$\mathbf{f(C)}$ **(PS model)**	$\mathbf{g(C)}$ **(Outcome model)**
6	$.5C_1 + .5C_2 + .1C_3$	$C_3 + C_4 + C_5 + \sum_{ij} C_i C_j$
7	$C_1 + C_2 + C_5$	$\sum_{ij} 0.5 C_i C_j$
8	$.2C_1 + .2C_2 + .2C_5$	$.25C_3 + (C_1 + C_2)^2 - (C_1^2 - C_3)^2 + (C_4^2 - .5C_5)(C_3 - .5C_4)$
9	$C_3 + C_4 + C_5 + \sum_{ij} C_i C_j$	$.5C_1 + .5C_2 + .1C_3$
10	$(C_1 + C_2 + .5C_3)^2$	$.5C_1 + .5C_3 + .5C_4$

Effects of potential confounders are allowed to be non-linear on both the treatment and outcome. Data is generated as follows: (1) $C_1, ..., C_5 \overset{iid}{\sim} N(0, 1)$; (2) $X \sim$ Bernoulli$(p = \text{expit}(f(C))$; and (3) $Y \sim \text{N}(\beta X + g(C), 1)$, where $f(\cdot)$ and $g(\cdot)$ are polynomial functions of C.

Table 3.1 when compared to $\widehat{\Delta}_{DR}^{MS}$, for each simulation scenario when the sample size is 500. Strikingly, we observe the smallest mean squared error using $\widehat{\Delta}_{DR}^{MA-dII}$ across all simulation scenarios presented here.

Table 3.4: Results of simulation study comparing double robust estimators for the average causal effect

	Percent reduction in MSE		
Scenario	$\widehat{\mathbf{\Delta}}_{\mathbf{DR}}^{\mathbf{MA-i}}$	$\widehat{\mathbf{\Delta}}_{\mathbf{DR}}^{\mathbf{MA-d}}$	$\widehat{\mathbf{\Delta}}_{\mathbf{DR}}^{\mathbf{MA-dII}}$
1	0.60	1.35	**5.42**
2	-0.03	-0.01	**5.23**
3	1.39	2.00	**59.33**
4	0.36	0.36	0.36
5	0.76	1.08	**1.27**
6	-1.16	-5.75	**29.1**
7	-12.5	-61.78	**90.1**
8	0.87	1.48	**29.9**
9	1.24	-0.65	**5.94**
10	0.19	0.47	**0.38**

The percent reduction in mean square error as compared to the model selected double robust estimator when the sample size is 500 for various model averaged double robust estimators. See Table 3.1 for definition of each estimator and Tables 3.2 and 3.3 for descriptions of each scenario. Bold indicates estimator with smallest MSE.

We have found that utilizing model averaging strategies on double robust estimators can reduce mean squared error as compared to the model selected double robust estimator. Our simulations support this claim, as at least one of the model averaged estimators always has a smaller mean squared error when compared with the model selected double robust estimator. This holds even when the true model's functional form is not included in the model class considered.

In the Group 1 simulations, where all effects are linear in the potential confounders, model averaging assuming prior model independence ,$\widehat{\Delta}_{DR}^{MA-i}$, reduces mean squared error up to 1.39% compared to model selection. This is a very modest gain, but demonstrates that simply applying model averaging to account for model uncertainty has a benefit

over using standard model selection procedures. The estimators $\widehat{\Delta}_{DR}^{MA-d}$ and $\widehat{\Delta}_{DR}^{MA-dII}$, where we assume prior model dependence, have reductions in mean squared error that are generally larger than those of $\widehat{\Delta}_{DR}^{MA-i}$. In fact, the estimator $\widehat{\Delta}_{DR}^{MA-dII}$ has reductions in mean squared error that range from 0.36% to 59.33%.

Considering Scenario 3, $\widehat{\Delta}_{DR}^{MA-dII}$ has 59.33% smaller mean squared error when compared to the model selected double robust estimator $\widehat{\Delta}_{DR}^{MS}$. Most of this reduction is in the variance of the estimator, as both estimators have little to no bias. This indicates that even in the case where all potential confounders are linear in both the propensity score and the outcome model, model averaging can reduce the variance of the double robust estimator dramatically if we assume prior model dependence and use the two-stage approach to cut model feedback.

We can explain this reduction in variance in Scenario 3 by noting that only C_1 and C_2 are confounders, while C_3, C_4, and C_5 are strongly associated with the exposure only. Therefore, using model selection on the propensity score model independently of the outcome model will tend to choose models that include C_3, C_4, and C_5. These three potential confounders are unrelated to the outcome, so their inclusion in the propensity score model only adds to the variance of the estimator. By utilizing model averaging with the prior model dependence given by Equation 3.7, we effectively restrict the model space of the propensity score a priori to be those models that include only the potential confounders that are associated with the outcome (C_1 and C_2). Thus, C_3, C_4, and C_5 are excluded from consideration by the prior distribution because they are unrelated to the outcome, and the reduction in mean squared error can be attributed to the correct identification of the C's associated with both the outcome and the treatment for use in the propensity score model. It is important to note here that the 59.33% reduction in mean squared error occurs when we have both assumed prior model dependence and used the two-stage approach for calculating model weights. The benefit of the latter point is argued in Section 3.2.3, but in this specific example, the posterior model probabilities used to construct $\widehat{\Delta}_{DR}^{MA-d}$ will favor the outcome model that includes all five potential confounders. This is ineffi-

cient because only C_1 and C_2 are associated with the outcome and including C_3, C_4, C_5 - which are only associated with X - into the outcome model, will lead to a large increase in the variance of the estimator. Therefore, cutting the feedback from the propensity score model into the outcome model when calculating the model weights improves efficiency.

In Scenario 4, the estimators $\widehat{\Delta}_{DR}^{MA-i}$, $\widehat{\Delta}_{DR}^{MA-d}$, and $\widehat{\Delta}_{DR}^{MA-dII}$ each reduce the mean squared error by 0.36% when compared with $\widehat{\Delta}_{DR}^{MS}$. This occurs because each of the five potential confounders are associated with both the exposure and the outcome, where those that are strongly associated with the outcome are moderately associated with the exposure and those that are strongly associated with the exposure are moderately associated with the outcome. Therefore, each method for estimating the posterior model weights will tend to select models that contain all five potential confounders.

Scenario 5 is a reproduction of Scenario 4, but with an additional 5 potential confounders that are unrelated to both the exposure and the outcome. In Scenarios 4 and 5, all the model averaged double robust estimators outperform the model selected double robust estimator. However, we note that when additional potential confounders are added (Scenario 5), the model averaged double robust estimators gain efficiency as compared to the model selected variety. This gain is expected to continue as more potential confounders are added, and since the estimator is scalable to a large number of potential confounders, the efficiency gain in using model averaging over model selection is likely to increase as the number of potential confounders grows.

Moving to Group 2 of the simulations, where the class of models considered is misspecified for either the propensity score or the outcome, $\widehat{\Delta}_{DR}^{MA-i}$ increases the mean squared error in 2 out of the 5 scenarios, the estimator $\widehat{\Delta}_{DR}^{MA-d}$ increases the mean square error in 3 out of 5 scenarios compared to $\widehat{\Delta}_{DR}^{MS}$, and no general conclusion about the comparison of model selection versus model averaging can be made. However, the estimator $\widehat{\Delta}_{DR}^{MA-dII}$ has a smaller mean squared error than $\widehat{\Delta}_{DR}^{MS}$ in all presented scenarios and reduces the mean squared error between 0.38% and 90.1%. In fact, $\widehat{\Delta}_{DR}^{MA-dII}$ appears to reduce mean

squared error the most when the outcome model has been misspecified (Scenarios 6-8). In all of the Scenarios 6-8, $\widehat{\Delta}_{DR}^{MA-dII}$ has at least 25% smaller mean squared error than $\widehat{\Delta}_{DR}^{MS}$.

In Scenario 7, the use of model averaging has reduced the variance (again, the mean squared error approximates the variance due to little to no bias) by 90.1%. To put this into perspective, if the model selected double robust estimator had a variance of 10, then the model averaged double robust estimator assuming prior model dependence and using the two-stage approach for calculating model weights would have a variance of about 1.

To illustrate the why cutting the feedback between the propensity score model and outcome model is effective, Tables 3.5 and 3.6 provide the marginal outcome model and propensity score model weights, respectively, used in the construction of $\widehat{\Delta}_{DR}^{MA-i}$, $\widehat{\Delta}_{DR}^{MA-d}$, and $\widehat{\Delta}_{DR}^{MA-dII}$ for Scenario 7 averaged over the 10,000 realizations. First, we will compare the model weights that are used in $\widehat{\Delta}_{DR}^{MA-i}$ and $\widehat{\Delta}_{DR}^{MA-dII}$ to describe why $\widehat{\Delta}_{DR}^{MA-i}$ increases the mean squared error by 12.5% while $\widehat{\Delta}_{DR}^{MA-dII}$ reduces mean square error by 90.1% when compared to the model selected double robust estimator. Note that the marginal outcome model weights used in $\widehat{\Delta}_{DR}^{MA-i}$ and $\widehat{\Delta}_{DR}^{MA-dII}$ are the same, so that the difference in the two estimators is due to the propensity score model weights. Referring to Table 3.6, the model weights used in construction of $\widehat{\Delta}_{DR}^{MA-i}$ assign 83.5% of the mass to the true propensity score model, while the weights used in construction of $\widehat{\Delta}_{DR}^{MA-dII}$ place the mass across many different propensity score models. The reduction in mean squared error can be attributed to the fact that when the outcome is non-linear in the potential confounders, it is unclear if adjusting for confounders linearly in the propensity score is optimal. The estimator $\widehat{\Delta}_{DR}^{MA-dII}$ captures this uncertainty, and the resulting model averaged double robust estimator averages over many different propensity score models resulting in a 90.1% reduction in mean squared error.

Next, we compare the model weights that are used in $\widehat{\Delta}_{DR}^{MA-d}$ and $\widehat{\Delta}_{DR}^{MA-dII}$ to describe why $\widehat{\Delta}_{DR}^{MA-d}$ increases the mean squared error by 61.78% while $\widehat{\Delta}_{DR}^{MA-dII}$ reduces mean square error by 90.1% when compared to $\widehat{\Delta}_{DR}^{MS}$. The prior model dependence forces a

Table 3.5: Marginal posterior outcome model weights in Scenario 7

Model	**Systematic Component**	$\widehat{\boldsymbol{\Delta}}_{\mathbf{DR}}^{\mathbf{MA-i}}$	$\widehat{\boldsymbol{\Delta}}_{\mathbf{DR}}^{\mathbf{MA-d}}$	$\widehat{\boldsymbol{\Delta}}_{\mathbf{DR}}^{\mathbf{MA-dII}}$
1	C_1	0.043	0	0.043
2	C_2	0.043	0	0.043
3	C_3	0.045	0	0.045
4	C_4	0.046	0	0.046
5	C_5	0.043	0	0.043
6	$C_1 + C_2$	0.031	0	0.031
7	$C_1 + C_3$	0.032	0	0.032
8	$C_1 + C_4$	0.032	0	0.032
9	$C_1 + C_5$	0.031	0	0.031
10	$C_2 + C_3$	0.032	0	0.032
11	$C_2 + C_4$	0.032	0	0.032
12	$C_2 + C_5$	0.032	0	0.032
13	$C_3 + C_4$	0.035	0	0.035
14	$C_3 + C_5$	0.032	0	0.032
15	$C_4 + C_5$	0.032	0	0.032
16	$C_1 + C_2 + C_3$	0.025	0	0.025
17	$C_1 + C_2 + C_4$	0.025	0	0.025
18	$C_1 + C_2 + C_5$	0.027	0.437	0.027
19	$C_1 + C_3 + C_4$	0.026	0	0.026
20	$C_1 + C_3 + C_5$	0.025	0	0.025
21	$C_1 + C_4 + C_5$	0.025	0	0.025
22	$C_2 + C_3 + C_4$	0.026	0	0.026
23	$C_2 + C_3 + C_5$	0.024	0	0.024
24	$C_2 + C_4 + C_5$	0.025	0	0.025
25	$C_3 + C_4 + C_5$	0.025	0	0.025
26	$C_1 + C_2 + C_3 + C_4$	0.022	0	0.022
27	$C_1 + C_2 + C_3 + C_5$	0.025	0.216	0.025
28	$C_1 + C_2 + C_4 + C_5$	0.025	0.216	0.025
29	$C_1 + C_3 + C_4 + C_5$	0.022	0	0.022
30	$C_2 + C_3 + C_4 + C_5$	0.022	0	0.022
31	$C_1 + C_2 + C_3 + C_4 + C_5$	0.027	0.131	0.027
32	intercept only	0.063	0	0.063

The marginal posterior outcome model weights used in the construction of $\widehat{\Delta}_{DR}^{MA-i}$, $\widehat{\Delta}_{DR}^{MA-d}$, and $\widehat{\Delta}_{DR}^{MA-dII}$ for each model in $\mathcal{M}^{om}$ in Scenario 7 and sample size 500 averaged over the 10,000 realizations. See Table 3.1 for a description of the estimators and Table 3.3 for a description of Scenario 7

Table 3.6: Marginal posterior propensity score model weights in Scenario 7

Model	**Systematic Component**	$\widehat{\mathbf{\Delta}}_{\mathbf{DR}}^{\mathbf{MA-i}}$	$\widehat{\mathbf{\Delta}}_{\mathbf{DR}}^{\mathbf{MA-d}}$	$\widehat{\mathbf{\Delta}}_{\mathbf{DR}}^{\mathbf{MA-dII}}$
1	C_1	0	0	0.126
2	C_2	0	0	0.127
3	C_3	0	0	0.021
4	C_4	0	0	0.021
5	C_5	0	0	0.126
6	$C_1 + C_2$	0	0	0.099
7	$C_1 + C_3$	0	0	0.003
8	$C_1 + C_4$	0	0	0.003
9	$C_1 + C_5$	0	0	0.094
10	$C_2 + C_3$	0	0	0.004
11	$C_2 + C_4$	0	0	0.002
12	$C_2 + C_5$	0	0	0.097
13	$C_3 + C_4$	0	0	0
14	$C_3 + C_5$	0	0	0.003
15	$C_4 + C_5$	0	0	0.003
16	$C_1 + C_2 + C_3$	0	0	0.003
17	$C_1 + C_2 + C_4$	0	0	0.003
18	$C_1 + C_2 + C_5$	0.835	0.941	0.095
19	$C_1 + C_3 + C_4$	0	0	0
20	$C_1 + C_3 + C_5$	0	0	0.006
21	$C_1 + C_4 + C_5$	0	0	0.003
22	$C_2 + C_3 + C_4$	0	0	0
23	$C_2 + C_3 + C_5$	0	0	0.003
24	$C_2 + C_4 + C_5$	0	0	0.003
25	$C_3 + C_4 + C_5$	0	0	0
26	$C_1 + C_2 + C_3 + C_4$	0	0	0
27	$C_1 + C_2 + C_3 + C_5$	0.08	0.029	0.004
28	$C_1 + C_2 + C_4 + C_5$	0.078	0.029	0.004
29	$C_1 + C_3 + C_4 + C_5$	0	0	0
30	$C_2 + C_3 + C_4 + C_5$	0	0	0
31	$C_1 + C_2 + C_3 + C_4 + C_5$	0.007	0.001	0
32	intercept only	0	0	0.146

The marginal posterior propensity score model weights used in the construction of $\widehat{\Delta}_{DR}^{MA-i}$, $\widehat{\Delta}_{DR}^{MA-d}$, and $\widehat{\Delta}_{DR}^{MA-dII}$ for each model in $\mathcal{M}^{om}$ in Scenario 7 and sample size 500 averaged over the 10,000 realizations. See Table 3.1 for a description of the estimators and Table 3.3 for a description of Scenario 7.

potential confounder that is included in the propensity score model to be included in the outcome model, and we see this through the marginal outcome model posterior probabilities used in the construction of $\widehat{\Delta}_{DR}^{MA-d}$. All of the mass is assigned to outcome models that include the three potential confounders that are associated with the treatment (C_1, C_2, and C_5) – only 4 of the 32 models have non-zero mass. The estimator $\widehat{\Delta}_{DR}^{MA-dII}$ distributes the outcome model weight more evenly across the model space, with all 32 outcome models receiving mass between 0.022 and 0.063. A similar result is observed in the marginal propensity score model weights, with $\widehat{\Delta}_{DR}^{MA-d}$ assigning 94.1% of the mass to the true propensity score model and $\widehat{\Delta}_{DR}^{MA-dII}$ distributing the weight across the model space. Therefore, the weights used in the estimator $\widehat{\Delta}_{DR}^{MA-d}$ tends to favor the potential confounders that are associated with the treatment in both the propensity score model and the outcome model. The data generating mechanism is non-linear in the potential confounders; using model selection or assigning most of the posterior weight to a few models that adjust for confounding linearly led to an inefficient estimate. This is important, as it emphasizes that model averaging provides the most gain in efficiency when there is non-linearities in the data generating mechanism.

To further emphasize this point, Figure 3.1 plots the model specific double robust estimators $\widehat{\Delta}_{ij}^{DR}$ versus their corresponding posterior weights p_{ij} used in the construction of $\widehat{\Delta}_{DR}^{MA-i}$, $\widehat{\Delta}_{DR}^{MA-d}$, and $\widehat{\Delta}_{DR}^{MA-dII}$ for a single realization of the data in Scenario 7. The vertical line is placed at the value of the corresponding model averaged estimator. It can be seen that when estimating both $\widehat{\Delta}_{DR}^{MA-i}$ and $\widehat{\Delta}_{DR}^{MA-d}$, the posterior mass is assigned to a few models that provide biased estimates of the true $\Delta = 1$. When estimating $\widehat{\Delta}_{DR}^{MA-dII}$, the posterior weight is spread across a different set of models that all provide a less biased estimate of Δ=1. As a reference, the model selected double robust estimate of Δ is $\widehat{\Delta}_{DR}^{MS} = 3.84$, which lies in the region of models that are assigned positive mass when estimating both $\widehat{\Delta}_{DR}^{MA-i}$ and $\widehat{\Delta}_{DR}^{MA-d}$.

Figure 3.1 provides results for a single realization of the data in Scenario 7, and as such, could be an artifact of randomness. To provide evidence of the contrary, Figure 3.2 pro-

vides a plot averaged over 10,000 realizations of the data in Scenario 7 that is constructed as follows: (1) for each simulated dataset, we round the model specific estimates $\widehat{\Delta}_{ij}^{DR}$ to the nearest whole number; (2) we assign each integer to the sum of the weights p_{ij} of the model specific double robust estimators that are mapped to that integer; and (3) we average the weights that are assigned to each integer over the 10,000 realizations of the data. In the estimation of $\widehat{\Delta}_{DR}^{MA-dII}$, approximately 80% of the posterior weight is assigned to models whose estimates round to the true value of $\Delta = 1$, while in the estimation of $\widehat{\Delta}_{DR}^{MA-i}$ and $\widehat{\Delta}_{DR}^{MA-d}$, only between 40% and 60% of the posterior mass is assigned to these same models. The model specific double robust estimators were rounded to the nearest integer to collapse the estimators based on the quality of the estimate within a given dataset. This allows us to summarize on average, how well do models that are assigned positive weight estimate $\Delta = 1$.

Putting the information from Figure 3.1 and 3.2 together, $\widehat{\Delta}_{DR}^{MA-dII}$ is a weighted average of model specific estimates that assigns most of the posterior weight to models that provide better estimates of $\Delta = 1$. This leads directly to $\widehat{\Delta}_{DR}^{MA-dII}$ reducing the mean squared error by 90.1% when compared to $\widehat{\Delta}_{DR}^{MS}$. In comparison, the estimators $\widehat{\Delta}_{DR}^{MA-i}$ and $\widehat{\Delta}_{DR}^{MA-d}$ fail to assign high posterior weight to models that provide good estimates of $\Delta = 1$, leading to more variable estimators.

The decision to compare the efficiency of the model averaged double robust estimator to that of the model selected double robust estimator was made because in practice, one must always make a decision regarding the models to be used. Without relying on expert knowledge, the only other alternative is to include all of the potential confounders in both the propensity score and outcome models. A sensitivity analysis was performed that indicates the results of our simulations are not sensitive to the choice of using model selection in place of the kitchen sink approach. Additionally, if we allow the potential confounders to be generated in a non-i.i.d. fashion, similar results hold.

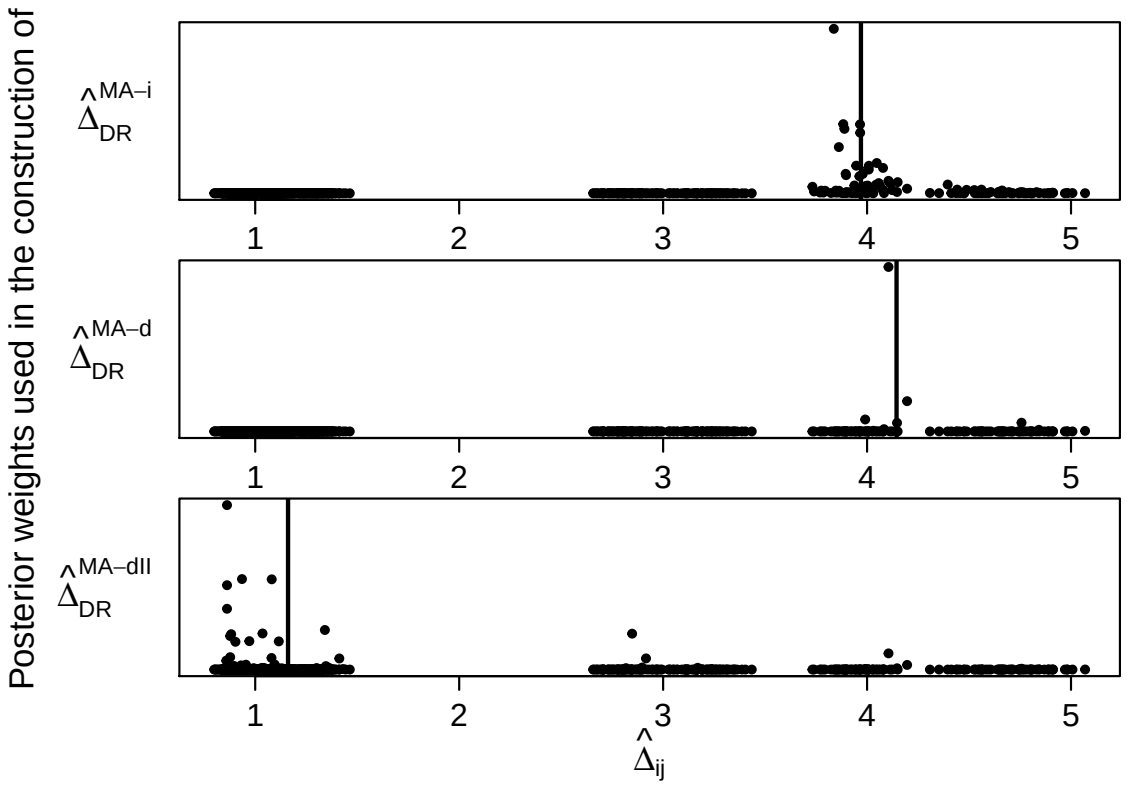

Figure 3.1: The model specific double robust estimators $\widehat{\Delta}^{DR}_{ij}$ versus their corresponding posterior weights p_{ij} used in the construction $\widehat{\Delta}^{MA-i}_{DR}$, $\widehat{\Delta}^{MA-d}_{DR}$, and $\widehat{\Delta}^{MA-dII}_{DR}$ of for a single realization of the data in Scenario 7. The vertical line is placed at the value of the corresponding model averaged estimator. The true value of Δ is 1. See Table 3.1 for definition of each estimator and Table 3.3 for a description of Scenario 7.

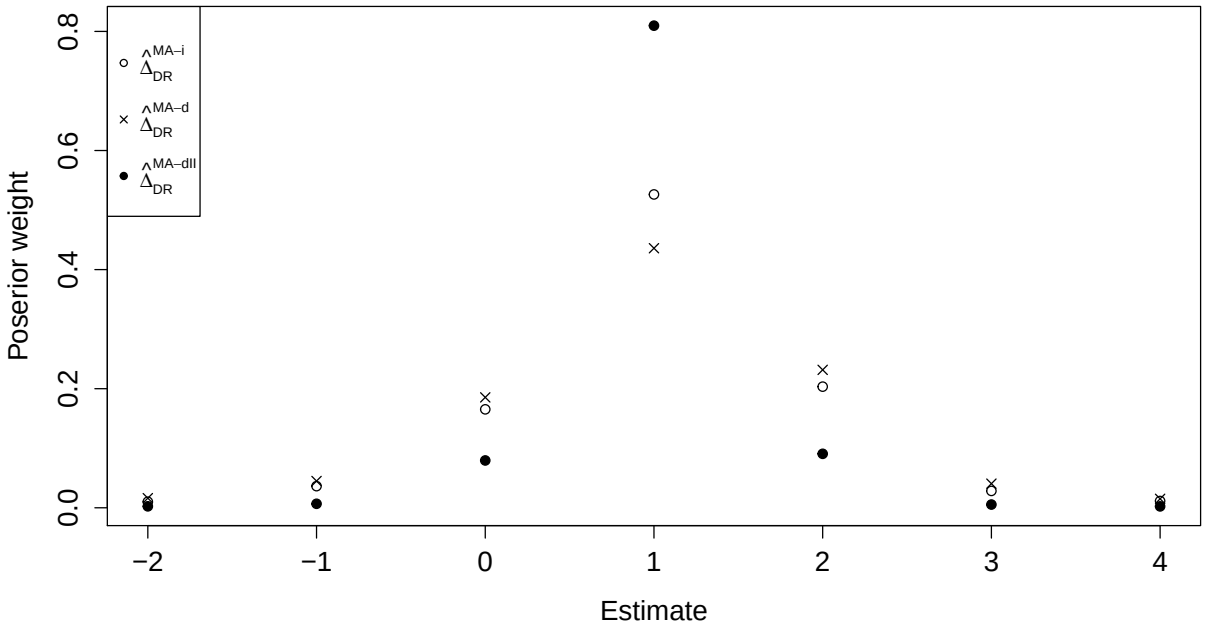

Figure 3.2: Summary of the posterior weights p_{ij} averaged over 10,000 realizations of the data in Scenario 7 that is constructed as follows: (1) for each simulated dataset, the model specific estimates $\widehat{\Delta}_{ij}^{DR}$ are rounded to the nearest whole number; (2) each integer is assigned the sum of the weights p_{ij} of the model specific double robust estimators that are mapped to that integer; and (3) average the weights that are assigned to each integer over the 10,000 realizations of the data. The horizontal axis is the value of the model specific double robust estimators that have been rounded to the nearest integer, and the vertical axis is the sum of the posterior weights of the corresponding model averaged double robust estimators that round to the specified integer averaged over the 10,000 realizations. The true value of Δ is 1. See Table 3.1 for definition of each estimator and Table 3.3 for a description of Scenario 7.

3.4 Discussion

We introduced model averaged double robust estimators, a new class of estimators for the average causal effect that account for model uncertainty. We proved that these estimators extend the popular double robustness property, by only requiring that the propensity score model or the outcome model be within a class of models. We also assessed small sample behavior: in realistic scenarios we showed substantial improvements over approaches that do not consider uncertainty in variable selection.

Our asymptotic results build on the most basic double robust estimator for the average causal effect. It has been demonstrated elsewhere that this double robust estimator can be biased especially when some of the estimated propensity scores are close to zero or are highly variable, and several adjustments to the estimator have been proposed (see (Robins et al., 2007; Cao et al., 2009; Tan, 2010) for discussion on this topic). The results on consistency of the model averaged double robust estimator will carry over to these other double robust estimators. Also, in the definition of the model averaged double robust estimator, we have assumed that the confounders' effect on the potential outcomes are the same between treatment groups, but this assumption is unnecessary. The methods presented in this paper can easily be extended to situations where the response surface differs between potential outcomes (Wang et al., 2012b) by using separate models for the potential outcomes, and independently calculating the posterior model probabilities.

We devised a two-stage approach for calculating the weights of the propensity score and outcome models. This two-stage approach utilizes a prior distribution on the model space that assumes dependence between a confounder's inclusion in the propensity score and the outcome model while cutting feedback from the propensity score model into the outcome model. Different choices of priors on the model space could have induced other desirable dependencies between the propensity score and the outcome model. For example, a similar dependence is implicit in the recent method proposed by Wang et al. (Wang et al., 2012a), for continuous exposures in the context of linear regression. This work has

recently been extended to binary exposures by Zigler and Dominici (Zigler and Dominici, 2012) in the context of stochastic search variable selections for propensity score models.

In our simulations, we have shown that through this two-stage approach for the calculation of model weights, one can reduce the mean squared error of the proposed estimator substantially —more than 90% in the most extreme cases. Reductions in mean squared error are largest in the likely situation when the outcome model is outside the model class considered and guaranteed to be misspecified. These results are not surprising, as there is a growing body of evidence that the use of non-informative priors for model selection in causal inference is not optimal (Brookhart, 2006; Crainiceanu et al., 2008; Schneeweiss et al., 2009; Vansteelandt et al., 2010; Wang et al., 2012a). While $\widehat{\Delta}_{DR}^{MA-dII}$ had the smallest mean squared error in the majority of the sensitivity analyses presented, in a few cases another estimator was more efficient. In these situations, the difference between the most efficient estimator and $\widehat{\Delta}_{DR}^{MA-dII}$ was minimal. It is unlikely that a researcher will correctly model the outcome; therefore, if a researcher chooses to use a doubly robust approach for estimation of the average causal effect, the two-stage model averaged double robust estimator with prior model dependence $\widehat{\Delta}_{DR}^{MA-dII}$ provides a very attractive implementation.

We restricted our class of models to be linear in the potential confounders, but even in the presence of non-linearities in the data generating mechanism, there were observed reductions in mean squared error as compared with the double robust estimator that performs model selection for the propensity score and the outcome model separately. Extension to nonlinear model classes would be conceptually straightforward.

Further work is needed to investigate whether these conclusions continue to hold when the set of potential confounders is larger and when the sample size is smaller (large p and small n). However, it is legitimate to conjecture that the improvements in efficiency should be greater in both directions, as both would emphasize the difference between model selection and model averaging. From this perspective, we expect that the gains presented in Section 3.3 should be conservative. For large model spaces, it is not feasi-

ble to explore every model combination as we did in our simulation study. However, one could implement Bayesian methods designed for model selection in the high dimensional data setting (George and McCulloch, 1993; O'Hara and Sillanpää, 2009; Johnson and Rossell, 2012), and use the corresponding posterior model weights in a model averaged double robust estimator.

The methods described in this paper share some similarities with the targeted maximum likelihood super-learner of van der Laan and colleagues (van der Laan et al., 2007; van der Laan, 2010). The super-learner acknowledges that no single learner is optimal and attempts to combine learners in a fashion to minimize a loss function via cross-validation. In this sense, model averaging the double robust estimator achieves the same goal, but instead combines candidate estimators via their posterior model probabilities. To further distinguish the methods, one must recognize that in both cases a researcher needs to characterize some underlying part of the true data distribution (e.g. the propensity score), denoted $\mathcal{Q}$, to estimate the average causal effect. The super-learner attempts to find the best estimate of $\mathcal{Q}$ upfront, and then uses this estimate of $\mathcal{Q}$ to construct a single estimator of the average causal effect. In contrast, the model averaged double robust estimator constructs several estimates of the average causal effect based on different parametric models that fully characterize $\mathcal{Q}$, and then directly averages these model specific estimates based on the posterior support of each model.

Causal inference approaches are increasingly used to analyze large observational studies, such as administrative databases used in comparative effectiveness research or environmental epidemiology. In these applications, there seldom is a clear-cut way of determining a priori the precise set of confounders of scientific relevance. At the same time, improvements in computing speed and parallelization are creating the opportunity for a more systematic investigation of alternative specifications for confounding adjustment. In this scenario, the proposed model averaging strategy shows great promise as a data analysis tool to perform robust and consistent inferences with good small sample properties.

4. Bias inflation due to exposure prediction in environmental epidemiology

[1]Matthew Cefalu and [1]Francesca Dominici

[1]Department of Biostatistics, Harvard School of Public Health

Abstract

Current epidemiological methods for studying the health effects of air pollution rely on exposure prediction models to align the air pollution exposure values with the outcome of interest. Such prediction is necessary because ambient air pollution is measured at a set of fixed and spatially sparse monitors that do not cover the entire study region, and in general, do not align spatially with the outcome. Many air pollution prediction methods have been suggested, including the nearest neighbor approach, kriging, and land-use regression. In land-use regression, geographic covariates are used in a regression model to improve the local heterogeneity of the predicted exposure, but little consideration is made as to whether the land-use covariates are also spatially correlated with the outcome. In this paper, we introduce the concept of bias inflation due to exposure prediction of a confounded health effect estimate by simultaneously considering exposure prediction and confounding, and discuss its impact on air pollution epidemiology. We derive a closed form expression for the bias of a health effect estimate when using a predicted exposure that decomposes into the product of two pieces: the bias due to the lack of adjustment for confounding and the bias inflation factor due to predicting the exposure. Importantly, we show that bias inflation factor can be large even when the confounding bias is small; therefore, our results suggest that exposure prediction and confounding adjustment need to be considered simultaneously.

4.1 Introduction

In the past two decades, there has been a wealth of epidemiological research on the health effects of air pollution (see Dominici et al. (2003); Pope (2007); Breysse et al. (2012) for reviews of the literature). Most published studies have found significant associations between short-term and long-term exposure to ambient levels of air pollution and a wide range of adverse health outcomes.

Due to the spatial nature of air pollution monitoring networks, spatial misalignment between the exposure and outcome is very common in these studies of air pollution and health. This occurs because the air pollution measurements are obtained from fixed monitoring locations, while the outcome data is generally not available at the exact monitor locations. As such, the great majority of cohort studies are affected by some sort of misalignment between exposure and outcome.

The current approach to align exposure and outcome is to use observed air pollution measurements at the monitor locations to develop a statistical model for predicting air pollution levels that align with the outcome data. Many different methods can be employed to predict missing air pollution values, including nearest neighbor and kriging approaches (Oliver and Webster, 1990; Madsen et al., 2008). These approaches typically lead to predicted exposure values that are spatially smoother than the true underlying exposure. Recently, land-use regression (LUR) has garnered much attention because of its ability to improve local variation in the exposure prediction by incorporating land-use (geographic) covariates into the prediction model. Hoek et al. (2008) provides a review of LUR models, and see others for application of LUR in epidemiology (Henderson et al., 2007; Ross et al., 2007; Yanosky et al., 2008; Sahsuvaroglu et al., 2009; Neupane et al., 2010; Kloog et al., 2012a,b; Cesaroni et al., 2013).

Another issue that is prevalent in cohort studies of air pollution and health is spatial confounding, which arises due to the complex spatial dependencies that exist between air pollution, the health outcome of interest, and other covariates. A researcher will employ expert knowledge in an attempt to control any spatial confounding through the use of covariates that vary in space. Great care is taken to minimize the magnitude of the bias in the health effect estimate, although it is unlikely that the bias has been completely negated.

Sheppard et al. (2011) provides a discussion of both confounding and exposure measurement error in air pollution epidemiology, and points out that exposure assessment should

be evaluated in the context of health effect estimation. With effect estimation in mind, it is known that: (1) better exposure prediction (i.e. smaller prediction error) does not necessarily lead to better effect estimation (i.e. smaller mean squared error) (Szpiro et al., 2011a); and (2) confounding can lead to biased effect estimation (Pope III and Burnett, 2007). However, the current literature treats confounding and exposure prediction as two separate statistical issues. That is, methods that account for the measurement error in the predicted exposure often fail to acknowledge the existence of confounding, while methods designed to control confounding often fail to acknowledge that the exposure has been predicted.

In this paper, we introduce the concept of *bias inflation due to exposure prediction of a confounded health effect estimate* by simultaneously considering exposure prediction and confounding and discuss its impact in the context of epidemiological studies of air pollution and health. We show that if confounding has not been sufficiently accounted for in the health effect model and a predicted exposure is used in place of the true exposure, then the bias of the health effect estimate can be larger (in magnitude) than the bias due to confounding when using the true exposure. We derive a closed form expression for the bias of a health effect estimate when using a predicted exposure that decomposes into the product of two pieces: the bias due to the lack of adjustment for confounding and a bias inflation factor due to predicting the exposure. Therefore, exposure prediction and confounding adjustment must be considered simultaneously.

4.2 Bias inflation due to exposure prediction

Bias inflation due to exposure prediction of a confounded health effect estimate occurs when there exists bias due to the lack of adjustment for confounding and exposure prediction is necessary. Therefore, to begin the discussion of bias inflation, we first must define what is meant by bias due to the lack of adjustment for confounding.

Let $\mathbf{C}_i$ be a set of normally distributed covariates with mean $\boldsymbol{\mu}_c$ and covariance Σ_c, and assume that the outcome Y_i and the exposure X_i are generated under the following linear models:

$$Y_i = \beta_0 X_i + \mathbf{C}_i \gamma_0 + \epsilon_i^y \tag{4.1}$$

$$X_i = \mathbf{C}_i \alpha_0 + \epsilon_i^x \tag{4.2}$$

where ϵ_i^y and ϵ_i^x are independent, normally distributed, mean zero error terms with variances $\sigma^2_{y|xc}$ and $\sigma^2_{x|c}$. Suppose interest lies in the estimation of the linear exposure-outcome relationship β_0, conditional on the covariates $\mathbf{C}_i$. Here, and throughout, no restriction is placed on γ_0 or α_0, and individual components of the vectors are free to be 0.

We define bias due to the lack of adjustment for confounding as the bias in our estimation of β_0 that is due to failure to control for the covariates $\mathbf{C}_i$. That is, if one were to ignore $\mathbf{C}_i$ when fitting the outcome regression model and instead fit $Y_i = \beta X_i + \epsilon_i$, then the least squares estimate for β, call it $\widehat{\beta}_x$, is biased. We call this the bias due to the lack of adjustment for confounding and denote it as $bias(\widehat{\beta}_x) = \mathrm{E}[\widehat{\beta}_x - \beta_0]$.

Now suppose that the exposure and outcome are completely misaligned (that is, either the exposure or the outcome is observed for all i, but not both). Further, let $W_i = \mathbf{C}_i \alpha_0$ be the predicted exposure with α_0 known. Consider fitting the outcome regression model that uses the predicted exposure W_i in place of the true exposure X_i and fails to control for any confounding ($Y_i = \beta W_i + \epsilon_i$). The bias of the least squares estimator for β, call it $\widehat{\beta}_w$, is given by:

$$bias(\widehat{\beta}_w) = \mathrm{E}[\widehat{\beta}_w - \beta_0] = bias(\widehat{\beta}_x)\frac{\sigma_x^2}{\sigma_w^2} \tag{4.3}$$

where $\sigma_x^2 = \sigma_w^2 + \sigma^2_{x|c}$ and $\sigma_w^2 = \alpha_0^T \Sigma_c \alpha_0$ denote the variances of X and W, respectively. We call the second term of Equation 4.3 ($\frac{\sigma_x^2}{\sigma_w^2}$) the bias inflation factor, and note that it is

equal to the inverse of the population R^2 when using W to predict X. From an intuitive standpoint, we expect that the variation in the true exposure σ_x^2 will always be more than the variation in the predicted exposure σ_w^2, and hence, the bias inflation factor is always greater than 1 (i.e. the R^2 is always less than 1).

Notice that the bias of $\widehat{\beta}_w$ is the product of two pieces: (1) the bias due to lack of adjustment for confounding assuming that the true exposure is known ($bias(\widehat{\beta}_x)$); and (2) the bias inflation factor due exposure prediction ($\frac{\sigma_x^2}{\sigma_w^2}$). It is easy to see that $bias(\widehat{\beta}_x) = 0$ implies that $bias(\widehat{\beta}_w) = 0$; therefore, bias inflation due to exposure prediction should only an issue if there is some uncontrolled confounding. However, even in the presence of uncontrolled confounding, $bias(\widehat{\beta}_x) \neq 0$ implies $bias(\widehat{\beta}_w) \neq 0$.

The bias inflation factor decreases as R^2 increases and goes to 1 as the exposure model is able to predict the true exposure X more accurately. Note that the bias inflation factor can be large even if the bias due to lack of adjustment for confounding is small. It is tempting to suggest that in an attempt to obtain an unbiased estimate of the health effect, a researcher should build an exposure model that more accurately predicts the true exposure (a model with the largest R^2). However, the relationship is not that simple. In fact, the bias of the health effect estimate can either increase or decrease in magnitude if a subset of the confounders are used in the exposure prediction model (see Section A.3.2 for closed form results). We will illustrate this concept using a simulated cohort study of the association between long-term exposure to $PM_{2.5}$ and cardiovascular disease in the New England region.

The previous results can easily be extended to situations where: (1) the outcome, exposure, and confounders are not assumed to be normally distributed; (2) the exposure prediction model uses a subset of the **C**s as defined in Equation 4.2; and/or (3) the outcome model controls for a subset of the **C**s as defined in Equation 4.1. For (1), we replace expectations with convergence in probability and all results continue to hold. For (2) and (3), closed form expressions for the biases are available in Section A.3.2.

In air pollution epidemiology, it is of great concern that there may be unmeasured spatial confounding. A researcher will attempt to control spatial confounding through the use of covariates that vary in space and is hopeful that the magnitude of the bias is minimal. The previous results can also be extended to incorporate these situations. Without going into mathematical details, it can be shown that if: (1) there is unmeasured spatial confounding; and (2) covariates that vary in space are used to predict air pollution, then there exists the potential for bias inflation due to exposure prediction. It is a challenge to untangle the complex spatial dependencies between the health outcome, air pollution, the measured covariates, and the unmeasured spatial confounders, and as such, it will be difficult to begin to quantify the magnitude of bias inflation due to exposure prediction in such studies. However, the existence of this bias can be demonstrated mathematically and by simulation, and much greater care is needed when predicted exposure levels are used in air pollution epidemiology research.

4.3 Simulations

4.3.1 Set up

Through the introduction of the concept of bias inflation due to exposure prediction, we have provided theoretical evidence that an exposure prediction model chosen solely on its ability to predict the true exposure may not lead to a better health effect estimate. We now provide a simple simulated example that clearly shows better prediction (higher R^2) does not imply better effect estimation and illustrates bias inflation due to exposure prediction.

Consider a hypothetical cohort study of the association between long-term exposure to $PM_{2.5}$ and cardiovascular disease in the New England region. Assume we have the cardiovascular hospitalization rates over the study period for each of the 2165 zipcodes in New England, and we wish to assign each zipcode to the mean $PM_{2.5}$ level over the study

period as a measure of exposure. Of the 2165 zipcodes, 57 have air pollution monitors within their boundaries, and the exposure for these zipcodes can be measured directly as the mean monitor value during the study period. For the remaining 2108 zipcodes, we assume the exposure values are missing and need to be predicted.

Figure 4.1 provides a map of the 2165 zipcodes in New England, with the 57 $PM_{2.5}$ monitoring locations marked with an x. We observe that the $PM_{2.5}$ monitors are sparse in New England, and tend to cluster near major population centers. As such, the spatial heterogeneity in $PM_{2.5}$ across New England will be difficult to capture based solely on spatial location (i.e. latitude and longitude).

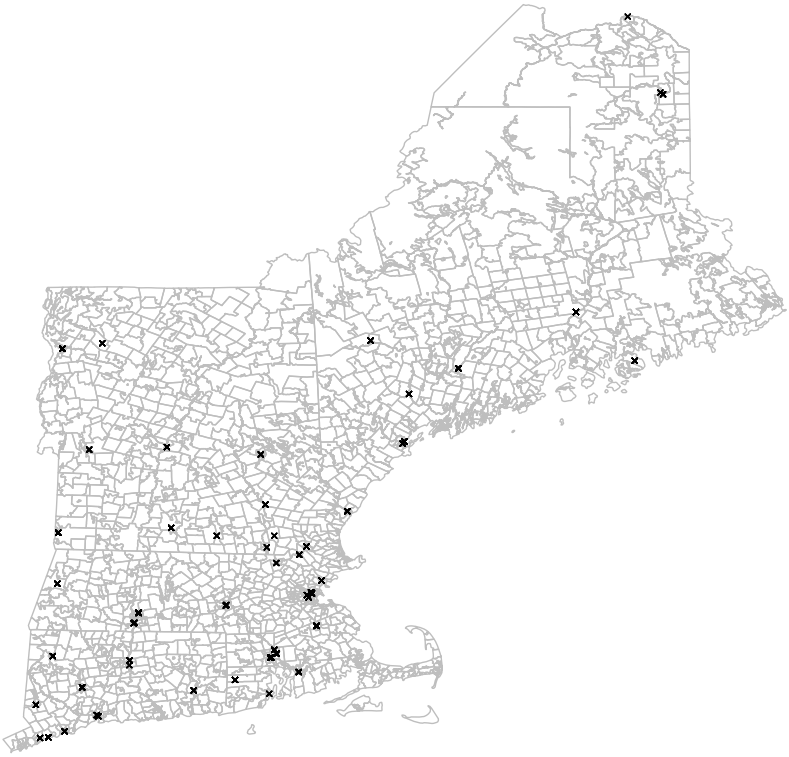

Figure 4.1: Map of the 2165 zip codes in New England, with the 57 $PM_{2.5}$ monitoring locations marked with an x

The intention of this simulation is to illustrate how the choice of covariates used in the $PM_{2.5}$ prediction model will affect the estimated health effect of long-term $PM_{2.5}$ exposure on hospitalization rates for cardiovascular disease. As such, we generate 1000 realizations of our hypothetical cohort in the following manner:

1. Use the observed distribution of 9 land-use (LU) covariates for each zipcode in New England. Table 4.1 provides a complete list and summary statistics for each land-use covariate considered.

2. Augment the 9 LU covariates with one $N(0,1)$ random variable, and denote the centered and standardized versions of these 10 covariates as $\mathbf{C}_i$.

3. Generate the exposure based on the relationship between the observed $PM_{2.5}$ levels and $\mathbf{C}$. That is, fit the exposure model $X_i = \mathbf{C}_i\alpha + \epsilon_x$ for the 57 zipcodes that have observed $PM_{2.5}$ measurements, and use the resulting $\widehat{\alpha}$ and $\widehat{\sigma}^2 = \widehat{var}(\epsilon_x)$ to generate a simulated "true" exposure as: $\widetilde{X}_i = \mathbf{C}_i\widehat{\alpha} + \mathcal{N}(0, \widehat{\sigma}^2)$

4. Generate the cardiovascular hospitalization rates: $ln(Y_i) = \beta_X\widetilde{X}_i + \mathbf{C}_i\gamma + \mathcal{N}(0, 0.467^2)$, where $\gamma = (0.01, 0.01, -0.1, -0.08, 0.8, -0.09, -0.09, 0.04, 0.008)$ and $\beta_X = 0.04$. Other choices of γ were considered and are available in the Section A.3.3.

5. Remove the "true" $PM_{2.5}$ values $\widetilde{X}$ from the dataset to reflect the zipcode that are missing exposure. The final dataset contains 57 zipcodes of $(Y_i, \widetilde{X}_i, \mathbf{C}_i)$ and 2108 zipcodes of $(Y_i, \mathbf{C}_i)$

The decision to not incorporate spatial correlation among the $PM_{2.5}$ values was to facilitate discussion, and it not reflective of what is expected in practice. This simulation scenario uses the worst case scenario; the same set of covariates that are used to predict and are also the ones that need to be used to adjust for confounding. In reality, there will be partial overlap between these two sets. See the Section A.3.2 for further discussion.

Table 4.1: Summary of 9 Land-use Covariates in New England

Covariate	**Minimum**	**1st quartile**	**Median**	**3rd quartile**	**Maximum**
% Forest	0.36	31.54	55.99	71.56	96.19
% Open space	2.71	64.47	89.07	94.91	100
% Urban	0	1.32	4.71	22.58	92.50
Traffic density	0	13.84	20.97	36.82	122.16
Elevation	10	40	120	240	1160
Distance to major road	0	0	2.91	13.22	133.87
Point emissions	0.001	0.001	0.0491	1.23	2015.46
Area emissions	0.003	0.260	1.19	8.16	86.01
Population per sq. km	0.13	25.71	105.18	363.71	4044.88

We will proceed using land-use regression (LUR) to estimate $PM_{2.5}$ levels that are missing from the study. However, since the decision was made to not incorporate spatial correlation among the $PM_{2.5}$ values in the simulated cohorts, our LUR regression will not involve spatial smoothing. Once the LUR is used to estimate the missing $PM_{2.5}$ values, an outcome regression is performed using a completed dataset that replaces the missing 2108 $PM_{2.5}$ values with their corresponding predicted values.

The only remaining decision for the purpose of our simulation is which LU covariates to include in the LUR. Considering every combination of the LU covariates would amount to $2^{10} = 1024$ possible models. Instead, we chose to consider 10 nested regression models that include the 10 LU covariates in order of their true predictive power of $PM_{2.5}$. The following summarizes the steps used to predict $PM_{2.5}$ and estimate the resulting health effect:

1. Fit the land-use regression model including only C_1 as a predictor for the 57 zip codes with observed $PM_{2.5}$
2. Estimate the 2108 missing $PM_{2.5}$ values, W, based on the model from Step 1
3. Estimate the effect of long-term $PM_{2.5}$ exposure on cardiovascular hospitalization rates using a regression model only including W as a predictor ($\ln(Y_i) = \beta W_i + \epsilon_i$)
4. Repeat 1-3, but using $\{C_1, C_2\}$, $\{C_1, C_2, C_3\}$, ... , $\{C_1, ..., C_{10}\}$ as predictors in the exposure regression model from Step 1

Note that in Step 3, we fit a regression model that fails to control confounding and gives a biased health effect estimate. The magnitude of this bias, which is given in closed form in Section A.3.2, is determined by a tradeoff between the bias due to lack of adjustment and the prediction accuracy of the $PM_{2.5}$ regression model and does not depend on the true value of β_X. As such, we consider only one value of $\beta_X = 0.04$.

4.3.2 Results

Figure 4.2 provides the R^2 from the LUR models and the corresponding bias of the health effect estimate from the hypothetical study of the association between long-term exposure to $PM_{2.5}$ and cardiovascular hospitalization rates in the New England region. The LUR that provides the health effect estimate with the smallest bias is the one that includes the first five LU covariates (% forrest, % open space, % urban, traffic density, and elevation) and has corresponding R^2 value of less than 0.6. By including the two additional covariates distance to major road and point emissions, the R^2 can be increased to 0.7, but results in a large bias. Of the 10 models considered, 5 have a smaller bias than the model that uses the true exposure (the dotted line), suggesting that a predicted exposure can either improve or worsen effect estimation when compared to the true exposure in the presence of uncontrolled confounding.

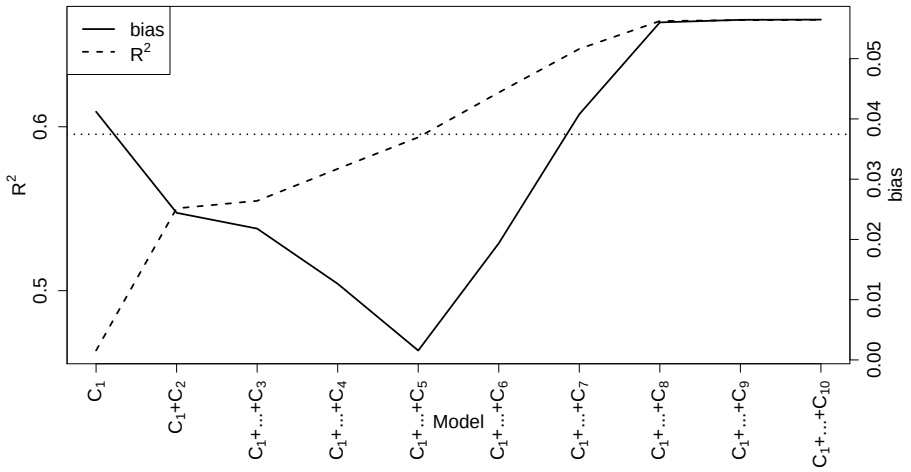

Figure 4.2: Tradeoff between R^2 and bias from the hypothetical cohort study of the association between long-term exposure to $PM_{2.5}$ and cardiovascular hospitalization rates in the New England region

Table 4.2 provides the percent of simulated datasets in which $H_0 : \beta = 0$ is rejected at the $\alpha = 0.05$ level when different subsets of C are used in the LUR to predict the exposure.

Table 4.2: Results of the hypothetical cohort study of the association between long-term exposure to $PM_{2.5}$ and cardiovascular disease in the New England region

Exp. Model	**% Reject H_0**	**% Reject H_0 & $\widehat{\beta} > 0$**	**% Reject H_0 & $\widehat{\beta} > 0$**	**%Bias**
$\{C_1\}$	100.0	100.0	0.0	102
$\{C_1, C_2\}$	99.8	99.8	0.0	60
$\{C_1, C_2, C_3\}$	99.8	99.8	0.0	53
$\{C_1, \cdots, C_4\}$	98.2	98.2	0.0	30
$\{C_1, \cdots, C_5\}$	89.6	89.6	0.0	3
$\{C_1, \cdots, C_6\}$	35.6	35.6	0.0	-50
$\{C_1, \cdots, C_7\}$	3.0	1.0	2.0	-104
$\{C_1, \cdots, C_8\}$	26.4	0.0	26.4	-142
$\{C_1, \cdots, C_9\}$	29.0	0.0	29.0	-143
$\{C_1, \cdots, C_{10}\}$	30.0	0.0	30.0	-144

Include is the percent of simulated datasets in which $H_0 : \beta = 0$ is rejected at the $\alpha = 0.05$ level when different subsets of **C** are used in the LUR to predict the exposure. Also included is the percent of simulations in which H_0 is rejected and $\widehat{\beta}$ is in the correct direction ($\widehat{\beta} > 0$), the percent of simulations in which H_0 is rejected and $\widehat{\beta}$ is in the wrong direction ($\widehat{\beta} < 0$), and the percent bias.

Also included is the percent of simulations in which H_0 is rejected and $\widehat{\beta}$ is in the correct direction ($\widehat{\beta} > 0$), the percent of simulations in which H_0 is rejected and $\widehat{\beta}$ is in the wrong direction ($\widehat{\beta} < 0$), and the percent bias.

As indicated in Figure 4.2, the exposure prediction model that minimizes the bias of the health effect estimate is $\mathcal{M}_5$, with corresponding bias of 3%. This LUR model rejects H_0 in 89.6% of the simulated datasets, with all rejection coming when the estimated health effect is in the correct direction. Therefore, if in this analysis we happened to choose $\mathcal{M}_5$ to predict $PM_{2.5}$, we would get nearly unbiased estimates of the effect of long-term $PM_{2.5}$ exposure on cardiovascular health and nearly 90% power to detect the true effect size.

However, if we deviate from this optimal model by either adding or removing LU covariates from the $PM_{2.5}$ prediction model, the percent bias ranges from -144% to 102%. By including the two additional covariates distance to major road and point emissions that increase R^2 by almost 0.1, we observe a bias of -104%, with H_0 being rejected only 3.0% of the time. Worse, two-thirds of the rejects occur when the estimated health effect is in the wrong direction. Thus, if in this analysis we happened to choose $\mathcal{M}_7$ to predict $PM_{2.5}$

, we would get biased results that estimate the health effect of long-term $PM_{2.5}$ exposure on cardiovascular hospitalization rates to be in the wrong direction.

This simple simulation illustrates that in the presence of uncontrolled confounding, a more accurate prediction of the exposure does not necessarily lead to a better health effect estimate. In fact, exposure prediction only exacerbates the problem of uncontrolled confounding, but all is not lost. Recall that in this hypothetical study, we purposefully fail to control for any confounding, but with a properly chosen $PM_{2.5}$ prediction model, we were able to return nearly unbiased effect estimates. In that situation, the bias inflation due to exposure prediction was beneficial for effect estimation. In general, we should be able to return valid effect estimates when using predicted exposure if: (1) confounding has sufficiently been controlled; or (2) an exposure prediction model is chosen to negate the effect of uncontrolled confounding.

The latter point is a challenging proposition, and current approaches in environmental epidemiology do not allow for consideration of the issue. In our simulation, we are able to determine which model should be used, but this is only because we know the true data generating mechanism. Without such knowledge, statistical methods for choosing an exposure prediction model to minimize the bias of the health effect estimate are needed.

4.4 Discussion

In this paper, we simultaneously consider spatial misalignment and spatial confounding in the context of cross-sectional studies, which rely almost entirely on the spatial variation between the exposure and the outcome to estimate the health effect of interest. We introduce the concept of *bias inflation due to exposure prediction of a confounded health effect estimate*, and purposely illustrate the point in a worst case (but not unrealistic scenario) where there is large overlap between covariates that are predictors of the exposure and covariates that are important measured confounders. We derive a closed form expres-

sion for the bias of a health effect estimate, and show that this bias is the product of two pieces: the bias due to the lack of adjustment for confounding and the bias inflation factor due to predicting the exposure with a set of measured covariates that are also measured confounders. Importantly, we show that bias inflation factor can be large even when the confounding bias is small; therefore, exposure prediction and confounding adjustment need to be considered simultaneously.

The potential for bias inflation due to exposure prediction can be demonstrated mathematically and by simulation, although quantifying the magnitude of the bias in practical applications will be conceptually challenging due to the complex spatial dependencies between the outcome, the exposure, the measured covariates, and the unmeasured spatial confounders. It is often the case that researchers build an air pollution prediction model that maximizes the spatial heterogeneity and cross-validated R^2, but do so independently of the outcome regression. We have provided evidence that such a process may lead to substantial bias inflation of the underlying health effect of interest.

Current statistical methods dealing with spatial misalignment and confounding adjustment treat the two topics as distinct issues. For example, methods to overcome spatial misalignment rely on exposure prediction, and exposure prediction can be viewed as a measurement error problem (Gryparis et al., 2009; Szpiro et al., 2011b). The measurement error can be decomposed into a Berkson-like component (Carroll et al., 1995) arising from modeling the exposure surface and a classical component arising from the estimation of the parameters of the exposure prediction model and several correction methods have been proposed (Gryparis et al., 2009; Szpiro et al., 2011b). However, these methods are only concerned with the bias of the health effect estimates due to measurement error and do not consider how predicting exposure with covariates that are correlated with the outcome might bias the health effect estimates. Similarly, methods designed for confounding adjustment do not acknowledge that the exposure has been predicted. For example, Wang et al. (2012a) was designed for the selection of confounders in the context of linear models for both the outcome and the exposure when the exposure has been fully observed.

Development of new statistical methods are needed that simultaneously predict exposure while adjusting for spatial confounding. The decision to include or exclude a potential confounder from either the outcome or the exposure model needs to be based on both the predictive power of the covariate on the exposure and the strength of the relationship with the outcome. An extension of Wang et al. (2012a) into the context of missing exposure could provide a foundation of methodologies used to simultaneously predict exposure and control confounding.

Our results do not address how spatial smoothing will affect the bias of a health effect in the presence of unmeasured spatial confounding. However, it is reasonable to postulate that bias inflation due to exposure prediction will exist when employing spatial smoothing. Such results would be related to the work of Dominici et al. (2004); they provide results to reduce confounding bias in the pollution-mortality relationship due to unmeasured time-varying factors such as season and influenza epidemics in the context of time series studies. One could adapt their results for use in cross-sectional studies of air pollution and health by indexing by space instead of time.

The issue of bias inflation due to exposure prediction was presented in the context of cross-sectional studies of air pollution and health. However, there is a likely statistical parallel for time series studies. If missing air pollution values are imputed using covariates that are temporally correlated with both air pollution and outcome, then a similar bias inflation is likely to occur. Meteorological covariates are one potential set of covariates that are temporally correlated with both air pollution and health.

The form of the bias inflation due to exposure prediction shares a remarkable similarity to that of a known result from causal inference; in the presence of unmeasured confounding, conditioning on instrumental variables can inflate the bias of the effect estimate (Bhattacharya and Vogt, 2012; Pearl, 2012).

The results of this paper assumed a simple linear relationship between the outcome, the exposure, and the confounders, but in practice, more complex models will be assumed

for both the exposure prediction model and the outcome regression model. However, even under these more complex models, there is potential for bias inflation of a health effect estimate due to exposure prediction, and much greater care is needed when using predicted exposure values in epidemiological studies of health.

A. Appendices

A.1 Efficient estimation of risk ratios from clustered binary data

A.1.1 Proof of results

Proof of Result 1: Recall that the nuisance tangent space is characterized by $\Lambda = \Lambda_{RM} + \Lambda_\alpha$, where Λ_{RM} is the nuisance tangent space from the restricted mean model and Λ_α is the closed linear space spanned by scores for α_0 along all regular parametric submodels. For any $A(X)\epsilon \in \Lambda_{RM}^{\perp}$, then

$$\begin{aligned}
\Pi\left[A(\mathbf{X})\epsilon|(\Lambda_{RM}+\Lambda_\alpha)^{\perp}\right] &= A(\mathbf{X})\epsilon - \Pi\left[A(\mathbf{X})\epsilon|\Lambda_{RM}+\Lambda_\alpha\right] \\
&= A(\mathbf{X})\epsilon - \Pi\left[A(\mathbf{X})\epsilon|\left\{\Lambda_\alpha - \Pi\left[\Lambda_\alpha|\Lambda_{RM}\right]\right\}\right] \\
&= A(\mathbf{X})\epsilon - \Pi\left[A(\mathbf{X})\epsilon|\Lambda_\alpha^*\right] \\
&= A(\mathbf{X})\epsilon - \frac{\mathrm{E}\left[A(\mathbf{X})\epsilon\epsilon^T V^{-1}(\mathbf{X})M(\mathbf{X})\mathbf{1}_k\right]}{\mathrm{E}\left[\mu^T(\mathbf{X})V^{-1}(\mathbf{X})\mu(\mathbf{X})\right]}\mu^T(\mathbf{X})V^{-1}(\mathbf{X})\epsilon \\
&= A(\mathbf{X})\epsilon - \frac{\mathrm{E}\left[A(\mathbf{X})\mu(\mathbf{X})\right]}{\mathrm{E}\left[\mu^T(\mathbf{X})V^{-1}(\mathbf{X})\mu(\mathbf{X})\right]}\mu^T(\mathbf{X})V^{-1}(\mathbf{X})\epsilon
\end{aligned}$$

where Λ_α^* is the closed linear space spanned by the efficient score for α_0 in $\mathcal{M}_{RM}$. Therefore, we have characterized the set of all influence functions for β_0 in the model $\mathcal{M}_{RM}$ that treats the baseline risk as a nuisance parameter as:

$$\Lambda_1^{\perp} = \left\{ \varphi(\mathbf{X}) = \mathrm{E}\left[A(\mathbf{X})D_\beta(\mathbf{X})\right]^{-1} A(\mathbf{X})\epsilon : \begin{array}{c} A(\mathbf{X}) = h(\mathbf{X}) - \frac{\mathrm{E}[h(\mathbf{X})\mu(\mathbf{X};\theta_0)]}{\mathrm{E}[\mu^T(\mathbf{X};\theta_0)V^{-1}(\mathbf{X})\mu(\mathbf{X};\theta_0)]}\mu^T(\mathbf{X};\theta_0)V^{-1}(\mathbf{X}), \\ h(\mathbf{X}) \text{ arbitrary} \end{array} \right\}$$

All that is left is to show $\Lambda^{\perp} = \Lambda_1^{\perp}$. For any $h(\mathbf{X}) \in \Lambda_1^{\perp}$, let $S(\mathbf{X}) = \left[h(\mathbf{X}) - \frac{\mathrm{E}[h(\mathbf{X})\mu^T(\mathbf{X})\mu(\mathbf{X})]}{\mathrm{E}[\mu^T(\mathbf{X})\mu(\mathbf{X})]}\right]\mu^T(\mathbf{X})$. Then,

$$\mathrm{E}\left[S(\mathbf{X})\mu(\mathbf{X})\right] = 0$$

so that $\Lambda_1^{\perp} \subset \Lambda^{\perp}$. Alternately, for any $S(\mathbf{X}) \in \Lambda^{\perp}$, let $h(\mathbf{X}) = S(\mathbf{X}) - \frac{\mathrm{E}[S(\mathbf{X})\mu(\mathbf{X})]}{\mathrm{E}[\mu^T(\mathbf{X})V^{-1}(\mathbf{X})\mu(\mathbf{X})]}\mu^T(\mathbf{X})V^{-1}(\mathbf{X})$. Then,

$$\mathrm{E}\left[h(\mathbf{X})\mu(\mathbf{X})\right] = 0$$

implying that $\Lambda^{\perp} \subset \Lambda_1^{\perp}$, and we are done.

Proof of Result 2: Let $U(h; \mathbf{X}, \alpha_0, \beta_0)$ be as defined in Result 1. Replace the log-baseline risk α_0 with an arbitrary value α. Then, for all h,

$$\begin{aligned}
\mathrm{E}[U(h;\mathbf{X},\alpha,\beta_0)] &= \mathrm{E}\left[h(\mathbf{X})\epsilon(\mathbf{X};\alpha,\beta_0) - \frac{\mathrm{E}\left[h(\mathbf{X})\mu(\mathbf{X};\alpha,\beta_0)\right]}{\mathrm{E}\left[\mu^T(\mathbf{X};\alpha,\beta_0)\mu(\mathbf{X};\alpha,\beta_0)\right]}\mu^T(\mathbf{X};\alpha,\beta_0)\epsilon(\mathbf{X};\alpha,\beta_0)\right] \\
&= \mathrm{E}\left[h(\mathbf{X})(Y - \mu(\mathbf{X};\alpha,\beta_0))\right] - \frac{\mathrm{E}\left[h(\mathbf{X})\mu(\mathbf{X};\alpha,\beta_0)\right]}{\mathrm{E}\left[\mu^T(\mathbf{X};\alpha,\beta_0)\mu(\mathbf{X};\alpha,\beta_0)\right]}\mathrm{E}\left[\mu^T(\mathbf{X};\alpha,\beta_0)(Y - \mu(\mathbf{X};\alpha,\beta_0))\right] \\
&= \mathrm{E}\left[h(\mathbf{X})\mathrm{E}[Y|\mathbf{X}]\right] - \mathrm{E}\left[h(\mathbf{X})\mu(\mathbf{X};\alpha,\beta_0))\right] - \frac{\mathrm{E}\left[h(\mathbf{X})e^{\mathbf{X}\beta_0}e^{\alpha}\right]}{\mathrm{E}\left[\mu^T(\mathbf{X};\alpha,\beta_0)\mu(\mathbf{X};\alpha,\beta_0)\right]}\mathrm{E}\left[\mu^T(\mathbf{X};\alpha,\beta_0)\mathrm{E}[Y|\mathbf{X}]\right] \\
&\quad + \frac{\mathrm{E}\left[h(\mathbf{X})\mu(\mathbf{X};\alpha,\beta_0)\right]}{\mathrm{E}\left[\mu^T(\mathbf{X};\alpha,\beta_0)\mu(\mathbf{X};\alpha,\beta_0)\right]}\mathrm{E}\left[\mu^T(\mathbf{X};\alpha,\beta_0)\mu(\mathbf{X};\alpha,\beta_0)\right] \\
&= \mathrm{E}\left[h(\mathbf{X})\mu(\mathbf{X};\alpha_0,\beta_0)\right] - \frac{\mathrm{E}\left[h(\mathbf{X})e^{\mathbf{X}\beta_0}e^{\alpha}\right]}{\mathrm{E}\left[\mu^T(\mathbf{X};\alpha,\beta_0)\mu(\mathbf{X};\alpha,\beta_0)\right]}\mathrm{E}\left[\mu^T(\mathbf{X};\alpha,\beta_0)\mu(\mathbf{X};\alpha_0,\beta_0)\right] \\
&= \mathrm{E}\left[h(\mathbf{X})\mu(\mathbf{X};\alpha_0,\beta_0)\right] - \frac{\mathrm{E}\left[h(\mathbf{X})e^{\mathbf{X}\beta_0}e^{\alpha_0}\right]}{\mathrm{E}\left[\mu^T(\mathbf{X};\alpha,\beta_0)\mu(\mathbf{X};\alpha,\beta_0)\right]}\mathrm{E}\left[\mu^T(\mathbf{X};\alpha,\beta_0)e^{\mathbf{X}\beta_0}e^{\alpha}\right] \\
&= 0
\end{aligned}$$

Proof of Result 3: Recall the efficient score is defined by $s_{\beta}^{eff} = \Pi[s_{\beta}|\Lambda^{\perp}]$, where s_{β} is the score for β_0. Under the restricted moment model, the efficient score (Bickel et al., 1998) for $\theta_0 = (\alpha_0, \beta_0)^T$ is given by:

$$s_{\theta}^{eff,RM} = (s_{\alpha}^{RM}, s_{\beta}^{RM})^T = \Pi\left[s_{\theta}|\Lambda_{RM}^{\perp}\right] = D^T(\mathbf{X})V^{-1}(X)\epsilon = (\mathbf{1}_k, \mathbf{X})^T M(\mathbf{X}|\theta_0)V^{-1}(\mathbf{X})\epsilon$$

where $D(\mathbf{X}) = \frac{\partial\mu(\mathbf{X}|\theta)}{\partial\theta^T}$, $M(\mathbf{X}|\theta) = \mathrm{diag}\left\{\mu(\mathbf{X}|\theta)\right\}$ is the (kxk) diagonal matrix made up of the elements of μ, and $V^{-1}(\mathbf{X}) = \mathrm{E}\left[\epsilon\epsilon^T\right]^{-1}$. Then, by definition of the efficient score and using arguments similar to Result 1:

$$s_\beta^{eff} = s_\beta^{RM} - \Pi\left[s_\beta^{RM} | \Lambda_\alpha^*\right]$$

where Λ_α^* is the closed linear space spanned by the efficient score for α_0 in $\mathcal{M}_{RM}$. Thus,

$$\begin{aligned}
s_\beta^{eff} &= s_\beta^* - \Pi\left[s_\beta^* \mid \Lambda_\alpha^*\right] \\
&= s_\beta^* - \mathrm{E}\left[s_\beta^* s_\alpha^{*T}\right] \mathrm{E}\left[s_\alpha^* s_\alpha^{*T}\right]^{-1} s_\alpha^* \\
&= \mathbf{X}^T M(\mathbf{X}|\alpha_0,\beta_0) V^{-1}(\mathbf{X})\epsilon - \mathrm{E}\left[\mathbf{X}^T M(\mathbf{X}|\alpha_0,\beta_0) V^{-1}(X)\epsilon\epsilon^T V^{-1}(\mathbf{X}) M^T(\mathbf{X}|\alpha_0,\beta_0)\mathbf{1}_k\right] \\
&\quad \mathrm{E}\left[\mathbf{1}_k^T M(\mathbf{X}|\alpha_0,\beta_0) V^{-1}(\mathbf{X})\epsilon\epsilon^T V^{-1}(\mathbf{X}) M^T(\mathbf{X}|\alpha_0,\beta_0)\mathbf{1}_k\right]^{-1} \mathbf{1}_k^T M(\mathbf{X}|\alpha_0,\beta_0) V^{-1}(\mathbf{X})\epsilon \\
&= \mathbf{X}^T M(\mathbf{X}|\alpha_0,\beta_0) V^{-1}(\mathbf{X})\epsilon - \mathrm{E}\left[\mathbf{X}^T M(\mathbf{X}|\alpha_0,\beta_0) V^{-1}(\mathbf{X}) M^T(\mathbf{X}|\alpha_0,\beta_0)\mathbf{1}_k\right] \\
&\quad \mathrm{E}\left[\mathbf{1}_k^T M(\mathbf{X}|\alpha_0,\beta_0) V^{-1}(\mathbf{X}) M^T(\mathbf{X}|\alpha_0,\beta_0)\mathbf{1}_k\right]^{-1} \mathbf{1}_k^T M(\mathbf{X}|\alpha_0,\beta_0) V^{-1}(\mathbf{X})\epsilon \\
&= \mathbf{X}^T M(\mathbf{X}|\alpha_0,\beta_0) V^{-1}(\mathbf{X})\epsilon - \mathrm{E}\left[\mathbf{X}^T M(\mathbf{X}|\alpha_0,\beta_0) V^{-1}(\mathbf{X})\mu(\mathbf{X}|\alpha_0,\beta_0)\right] \\
&\quad \mathrm{E}\left[\mu^T(\mathbf{X}|\alpha_0,\beta_0) V^{-1}(\mathbf{X})\mu(\mathbf{X}|\alpha_0,\beta_0)\right]^{-1} \mu^T(\mathbf{X}|\alpha_0,\beta_0) V^{-1}(\mathbf{X})\epsilon
\end{aligned}$$

A.1.2 Acknowledgements

We would like to thank Felton Earls and Mary Carlson for providing the *Young Citizens* data, and Lisa Yelland for the sharing of her code. Support for this research was provided by National Institute of Environmental Health Sciences grant 5T32ES007142 and NIH grants R01ES020337-01, R21ES019712, U54GM088558, and R0151164

A.2 Model averaged double robust estimation

A.2.1 Consistency under the dependent prior: A different view

The prior provided in the paper can be relaxed so that inclusion-exclusion criteria is not strict. In fact, it can written in a form that is close to the prior of Wang et al. (2012a). Let the class of model be defined by the indicators α^X and α^Y, where α^X is the indicator that a particular covariate is included in the propensity score model and α^Y is the indicator that a particular covariate is included in the outcome model. Using this notation, the propensity score model can be written as $g(\mathrm{E}[X|C]) = \xi_0 + \sum_{k=1}^{p} \alpha_k^X \xi_k C_k$ for some link function $g(\cdot)$, and the outcome model can be written as $\mathrm{E}[Y|X, C] = \gamma_0 + \beta X + \sum_{k=1}^{p} \alpha_k^Y \gamma_k C_k$. Relaxing the inclusion-exclusion criteria, the prior model dependence given by Equation 3.7 can be written as:

$$\begin{aligned}
\frac{P(\alpha^Y = 1|\alpha^X = 1)}{P(\alpha^Y = 0|\alpha^X = 1)} &= \omega \\
\frac{P(\alpha^X = 1|\alpha^Y = 1)}{P(\alpha^X = 0|\alpha^Y = 1)} &= 1 \\
P(\alpha^Y = 1) &= \frac{1}{2}
\end{aligned}$$

Note that this implies,

$$\begin{aligned}
P(\alpha^Y = 1, \alpha^X = 1) &= \frac{1}{4} \\
P(\alpha^Y = 1, \alpha^X = 0) &= \frac{1}{4} \\
P(\alpha^Y = 0, \alpha^X = 1) &= \frac{1}{4\omega} \\
P(\alpha^Y = 0, \alpha^X = 0) &= \frac{2\omega - 1}{4\omega}
\end{aligned}$$

So for any finite ω, the prior distribution does not affect the consistency of the posterior probabilities because the prior does not restrict model space, and the consistency of the posterior model probabilities relies on the consistency of the Bayes factor. In other words, because we have not restricted the model space, the likelihood will overpower the prior for large sample sizes.

With this in mind, the MA-DR estimator is consistent for any finite ω in the prior specification above. The prior presented in Equation 3.7 is for $\omega = \infty$, and we do not believe the consistency result will hold. However, the strict prior ($\omega = \infty$) leads to a posterior that is computationally much less burdensome than for any other choice of ω, except $\omega = 1$.

Thus, we view the $\omega = \infty$ case as an approximation to any large choice of ω. This is a reasonable approximation because for large ω, the prior will overwhelm the likelihood in finite samples. Therefore, the prior model dependence does not lead to a estimator that is consistent for the average causal effect, but is an approximation of an estimator that is consistent for the average causal effect.

A.2.2 Additional simulations

This set of simulations expands both the set of simulation scenarios along with the estimators being compared. Table A.1 provides a description of each estimator included in these simulations. A full description of all scenarios can be found in Table A.2 and Table A.3. All simulations set $\beta = 1$ and use a sample size of 500 with $10,000$ replications.

Table A.4 and Table A.5 provide the mean squared error and the bias of each estimator under each additional simulation scenario. These simulations highlight a few additional points that were not covered in the original paper. First, applying model averaging to only a parametric or IPW estimator does not perform as favorably as the model averaged double robust estimator. Specifically, consider $\widehat{\Delta}_{DR}^{MA-dII}$, $\widehat{\Delta}_{IPW}^{MA}$, and $\widehat{\Delta}_{para}^{MA}$.

First, looking at the MSE of $\widehat{\Delta}_{IPW}^{MA}$, it is considerably higher in many scenarios. Take

Table A.1: Description of all estimators used in the additional simulation study comparing estimators for the average causal effect

Estimator	Description
$\widehat{\Delta}^{MA}_{para}$	Model averaged parametric estimator
$\widehat{\Delta}^{MA}_{IPW}$	Model averaged IPW estimator
$\widehat{\Delta}^{MS}_{DR}$	Model selected double robust estimator that chooses propensity model and the outcome model based on the BIC
$\widehat{\Delta}^{MS-II}_{DR}$	Model selected double robust estimator that chooses the outcome model first, and then restricts the choice of the propensity score model to be a subset of the outcome model as in Equation 3.7
$\widehat{\Delta}^{MA-i}_{DR}$	MA-DR estimator assuming prior model independence
$\widehat{\Delta}^{MA-d}_{DR}$	MA-DR estimator assuming prior model dependence defined by Equation 3.7
$\widehat{\Delta}^{MA-dII}_{DR}$	MA-DR estimator assuming prior model dependence defined by Equation 3.7 and using the two-stage approach for calculating model weights
$\widehat{\Delta}^{MA-BAC}_{DR}$	MA-DR estimator assuming prior model dependence defined by Wang et al. (2012a)
$\widehat{\Delta}^{MA-\omega=10}_{DR}$	MA-DR estimator assuming prior model dependence with $\omega = 10$

Included is (1) the type of estimator; and (2) the choice of prior distribution for the model space. All Bayes factors are estimated using the BIC approximation.

Table A.2: Description of Group 1 in the additional simulation study comparing estimators for the average causal effect

Scenario	$\boldsymbol{\alpha^{\mathrm{ps}}}$ **(PS model)**	$\boldsymbol{\alpha^{\mathrm{om}}}$ **(Outcome model)**
1	(.1,.1,.01,0,0)	(.5,0,1,.5,0)
2	(.5,.5,.1,0,0)	(.5,0,1,.5,0)
3	(1,.5,.1,0,0)	(.5,1,2,1,0)
4	(.3,0,0,0,0)	(1,0,0,0,0)
5	(.4,.3,.2,.1,0)	(0,0,0,0,0)
6	(.5,.4,.3,.2,.1)	(.5,1,1.5,2,2.5)
7	(1,1,0,0,0)	(.2,.2,2,2,2)
8	(.05,.05,.5,.5,.5)	(2,2,.2,.2,.2)
9	(0.1,.025,.012,0.053,0.034)	(.5,.53,.22,.44,.62)
10	(0,0,0,0,0)	(1,0,0,0,0)
11	(.1,-.1,.01,0,0)	(-.5,0,1,.5,0)
12	(-.5,.5,.1,0,0)	(.5,0,1,-.5,0)
13	(1,-.5,-.1,0,0)	(.5,1,2,1,0)
14	(.3,0,0,0,0)	(-1,0,0,0,0)
15	(.4,-.3,-.2,.1,0)	(0,0,0,0,0)
16	(.5,.4,-.3,.2,.-1)	(.5,1,-1.5,2,-2.5)
17	(1,1,0,0,0)	(.2,-.2,2,2,2)
18	(.05,-.05,-.5,.5,.5)	(-2,2,.2,.2,.2)
19	(-0.1,.025,.012,-0.053,0.034)	(-.5,.53,.22,.44,-.62)
20	(0,0,0,0,0)	(-1,1,0,0,0)
21	(.1,.1,1,1,1)	(2,2,0,0,0)
22	(1,1,0,0,0)	(.5,.5,2,2,2)
23	(1,1,0,0,0)	(.8,.8,2,2,2)

All effects of confounders are linear on both the treatment and outcome. Data is generated as follows: (1) $C_1, ..., C_5 \overset{iid}{\sim} N(0,1)$; (2) $X \sim$ Bernoulli$(p = \mathrm{expit}(C\alpha^{ps}))$; and (3) $Y \sim \mathrm{N}(\beta X + C\alpha^{om}, 1)$

Table A.3: Description of Group 2 in the additional simulation study comparing estimators for the average causal effect

Scenario	$\mathbf{f(C)}$ **(PS model)**	$\mathbf{g(C)}$ **(Outcome model)**
24	$.5C_1 + .5C_2 + .1C_3$	$C_3 + C_4 + C_5 + \sum_{ij} C_i C_j$
25	$C_1 + C_2 + C_5$	$\sum_{ij} 0.5 C_i C_j$
26	$.2C_1 + .2C_2 + .2C_5$	$.25C_3 + (C_1 + C_2)^2 - (C_1^2 - C_3)^2 + (C_4^2 - .5C_5)(C_3 - .5C_4)$
27	$C_3 + C_4 + C_5 - \sum_{ij} C_i C_j$	$.5C_1 + .5C_2 + .1C_3$
28	$C_3 + C_4 + C_5 + \sum_{ij} C_i C_j$	$C_1 + C_2 + .5C_3$
29	$(C_1 + C_2 + .5C_3)^2$	$.5C_1 + .5C_3 + .5C_4$
30	$(C_1 + C_2 + .5C_3)^2$	$C_1 + C_3 + 2C_4$
31	$.5C_1 + .5C_2 + .1C_3$	$C_3 + C_4 + C_5 - \sum_{ij} C_i C_j$
32	$C_1 - C_2 - C_5$	$-\sum_{ij} 0.5 C_i C_j$
33	$(-C_1 + C_2 + .5C_3)^2$	$-.5C_1 + .5C_3 + .5C_4$
34	$(-C_1 + C_2 + .5C_3)^2$	$C_1 + C_3 - 2C_4$
35	$.2C_1 + .2C_2 + .2C_5$	$(1 + .5 * (C_1 + C_2 + C_3 + C_4 + C_5))^2$

Effects of potential confounders are allowed to be non-linear on both the treatment and outcome. Data is generated as follows: (1) $C_1, ..., C_5 \overset{iid}{\sim} N(0, 1)$; (2) $X \sim$ Bernoulli$(p = \text{expit}(f(C))$; and (3) $Y \sim \text{N}(\beta X + g(C), 1)$, where $f(\cdot)$ and $g(\cdot)$ are polynomial functions of C.

Table A.4: Mean square error (10^{-3}) from additional simulation study comparing estimators of the average causal effect

Scenario	$\widehat{\Delta}^{\mathrm{MA-i}}_{\mathrm{DR}}$	$\widehat{\Delta}^{\mathrm{MA}}_{\mathrm{para}}$	$\widehat{\Delta}^{\mathrm{MA}}_{\mathrm{IPW}}$	$\widehat{\Delta}_{\mathrm{DR}}$	$\widehat{\Delta}^{\mathrm{MA-d}}_{\mathrm{DR}}$	$\widehat{\Delta}^{\mathrm{MA-ii}}_{\mathrm{DR}}$	$\widehat{\Delta}^{\mathrm{MA-BAC}}_{\mathrm{DR}}$	$\widehat{\Delta}^{\mathrm{MA-}\omega=10}_{\mathrm{DR}}$	$\widehat{\Delta}^{\mathrm{MS-II}}_{\mathrm{DR}}$
1	8.08	8.08	14.42	8.09	8.07	8.07	8.08	8.07	8.07
2	9.63	9.37	20.97	9.61	9.62	9.31	9.63	9.62	9.28
3	12.03	32.77	83.21	12.04	12.03	12.02	12.05	12.03	12.04
4	8.27	8.34	10.52	8.25	8.24	8.25	8.26	8.24	8.24
5	8.22	7.72	8.25	8.31	8.15	7.88	8.21	8.16	7.84
6	9.73	19.79	268.51	9.75	9.73	9.73	9.75	9.73	9.75
7	14.81	57.88	147.66	14.78	14.79	14.7	14.81	14.79	14.89
8	10.76	15.84	85.18	10.78	10.76	10.78	10.76	10.76	10.8
9	7.82	7.81	15.7	7.82	7.82	7.82	7.82	7.82	7.82
10	8.89	8.88	12.73	8.9	8.88	8.89	8.9	8.88	8.91
11	8.5	8.49	14.49	8.53	8.47	8.49	8.49	8.48	8.51
12	9.44	9.25	19.32	9.43	9.44	9.17	9.45	9.44	9.32
13	11.62	21.04	72.78	11.67	11.61	11.62	11.63	11.61	11.66
14	8.58	8.6	10.2	8.57	8.57	8.57	8.58	8.58	8.56
15	8.68	8.31	8.7	8.73	8.64	8.45	8.69	8.64	8.47
16	9.33	19.76	261.22	9.34	9.33	9.33	9.36	9.33	9.34
17	16.05	61.21	156.57	16.04	16.03	15.86	16.05	16.03	16.11
18	10.48	16.54	87.34	10.47	10.48	10.51	10.48	10.48	10.49
19	7.09	7.08	11.59	7.1	7.09	7.09	7.09	7.09	7.1
20	8.55	8.52	15.97	8.56	8.54	8.54	8.55	8.54	8.56
21	23.87	11.48	248.54	24.24	23.52	9.77	23.61	23.57	9.89
22	14.65	76	164.79	14.67	14.65	14.65	14.64	14.65	14.67
23	14.98	94.27	174.21	14.98	14.98	14.98	14.99	14.98	14.98
24	665.95	394.17	668.58	656.11	701.39	478.6	705.12	690.74	455.9
25	16.86	24.87	30.04	16.93	17.14	16.34	17.54	17.03	16.23
26	1620.17	69.22	1804.57	1436.4	2434.01	178.84	2207.9	2192.64	241.8
27	10.17	10.27	13.55	10.24	10.13	10.16	10.18	10.14	10.17
28	11.13	12.8	40.13	11.13	11.11	11.12	11.14	11.11	11.12
29	660.05	389.61	654.94	651.78	691.82	480.4	696.07	682.42	461.88
30	16.15	27.33	34.5	16.46	16.5	15.3	16.87	16.37	15.33
31	375.27	132.11	424.67	382.23	396.73	192.75	389.9	391.2	243.65
32	10.82	10.98	14.13	10.85	10.79	10.79	10.81	10.79	10.8
33	10.83	11.16	36.42	10.88	10.82	10.82	10.84	10.82	10.84
34	800.77	704.92	886.87	807.11	796.17	770.58	807.09	796.63	772.05
35	39.36	37.4	113.62	39.43	39.36	39.36	39.79	39.36	39.43

Table A.5: Bias (10^{-3}) from additional simulation study comparing estimators of the average causal effect

Scenario	$\widehat{\Delta}^{\mathrm{MA-i}}_{\mathrm{DR}}$	$\widehat{\Delta}^{\mathrm{MA}}_{\mathrm{para}}$	$\widehat{\Delta}^{\mathrm{MA}}_{\mathrm{IPW}}$	$\widehat{\Delta}_{\mathrm{DR}}$	$\widehat{\Delta}^{\mathrm{MA-d}}_{\mathrm{DR}}$	$\widehat{\Delta}^{\mathrm{MA-ii}}_{\mathrm{DR}}$	$\widehat{\Delta}^{\mathrm{MA-BAC}}_{\mathrm{DR}}$	$\widehat{\Delta}^{\mathrm{MA-}\omega=10}_{\mathrm{DR}}$	$\widehat{\Delta}^{\mathrm{MS-II}}_{\mathrm{DR}}$
1	0.38	0.36	33.9	0.37	0.29	0.31	0.36	0.3	0.17
2	5.47	5.49	79.3	5.5	5.44	5.01	5.45	5.45	4.68
3	2.03	1.01	152.31	2.09	2.03	1.95	2	2.03	2
4	0.15	0.38	28.26	0.23	0.24	0.22	0.16	0.23	0.34
5	0.93	1.96	0.84	1.09	1.05	1.66	0.94	1.04	2.55
6	2.9	2.3	460.38	2.73	2.9	2.9	2.88	2.9	2.73
7	3.22	4.83	10.42	3.15	3.11	3.88	3.13	3.13	3.6
8	0.94	6.84	158.35	0.87	0.93	1.34	0.88	0.93	0.93
9	3.77	3.97	68.16	3.76	3.77	3.77	3.76	3.77	3.76
10	1.07	1.06	0.76	1.04	1.08	1.06	1.06	1.08	1.04
11	5.14	4.98	25.07	5.17	5.16	5.14	5.14	5.16	5.23
12	4.4	4.2	62.96	4.41	4.32	4.15	4.37	4.34	4.12
13	0.65	5.33	157.51	0.5	0.62	0.62	0.66	0.63	0.47
14	2.47	2.33	24.08	2.49	2.53	2.48	2.46	2.52	2.56
15	1.81	1.11	1.65	2.16	1.73	1.43	1.83	1.73	1.65
16	0.52	3.51	451.29	0.61	0.52	0.52	0.55	0.52	0.61
17	4.68	7.38	0.26	4.61	4.71	4.48	4.74	4.7	4.49
18	1.2	1.04	157.32	1.18	1.19	1.4	1.21	1.19	1.27
19	0.01	0.03	10.73	0	0.02	0.01	0.01	0.02	0
20	1.87	1.69	1.8	1.98	1.83	1.86	1.87	1.83	1.9
21	2.47	2.11	266.9	2.83	2.48	2.99	2.54	2.51	4.65
22	8.3	17.48	32.44	8.35	8.3	8.3	8.3	8.3	8.35
23	3.43	5.13	1.79	3.37	3.43	3.43	3.4	3.43	3.37
24	29.99	26.11	98.99	29.28	32.41	25.91	29.59	32.32	25.7
25	3.04	65.79	87.73	3.32	2.66	2.8	2.74	2.78	2.97
26	27	8.85	35.44	24.79	40.22	10.26	36.14	37.32	8.67
27	0.1	0.83	2.65	0.28	0.13	0.17	0.14	0.12	0.27
28	3.11	4.04	8.17	3.04	3.1	3.1	3.12	3.1	3.16
29	7.9	7.2	71.98	5.98	13.68	6.04	8.75	12.92	4.22
30	1.72	79.7	105.66	1.47	1.95	1.57	2.02	1.93	0.72
31	18.33	1	22.73	17.28	18.84	5.38	18.35	18.56	5.97
32	4.93	4.38	4.43	4.8	4.79	4.85	4.92	4.81	4.58
33	1	1.08	5.65	1.03	1.11	1.04	0.95	1.1	1.16
34	5.68	13.14	239.35	6.35	7.42	9.58	5.31	7.19	10.72
35	7.27	8.1	240.04	7.27	7.27	7.27	7.25	7.27	7.27

Scenario 6 for example, where all 5 covariates are moderate confounders. The MSE is $268.5\text{x}10^{-3}$, while all other estimators have MSE less than $20\text{x}10^{-3}$. Quickly taking a look at the bias of this estimator, we see a value of 0.460, corresponding to 46% bias. Therefore, the model averaged IPW estimator, assuming an independent prior on the model space, can provide highly variable and highly biased results.

Next, consider $\widehat{\Delta}_{para}^{MA}$. This estimator performs more favorably when compared with $\widehat{\Delta}_{DR}^{MA-dII}$, which is not all that surprising since in many of the simulations, the outcome model class is correctly specified. In 12 of the 35 simulation scenarios, $\widehat{\Delta}_{para}^{MA}$ has smaller mean squared error than $\widehat{\Delta}_{DR}^{MA-dII}$. However, when comparing the biases of the two estimators, we can point to several example where the bias of $\widehat{\Delta}_{para}^{MA}$ is considerably more than that of $\widehat{\Delta}_{DR}^{MA-dII}$. Considering only Scenario 25, the bias of $\widehat{\Delta}_{DR}^{MA-dII}$ is $65.79\text{x}10^{-3}$, while the bias of $\widehat{\Delta}_{DR}^{MA-dII}$ is only $2.8\text{x}10^{-3}$. This is approximately a 95% reduction in the bias. This is a situation where the data generating mechanism in the propensity score model is non-linear in the confounders, while the true outcome model is linear in the confounders. Therefore, even though we have a properly specified outcome model class, we cannot return a valid effect estimate due to the separation of the treatment groups.

We believe this verifies that simply using model averaging on either the IPW or a parametric estimator may lead to inefficient and/or biased effect estimates. Also included in this simulation is the MA-DR estimator that assume the prior of Wang et al. (2012a), and the estimator that assume the prior specified above with $\omega = 10$. These are not discussed in detail, but note that they behave similarly to $\widehat{\Delta}_{DR}^{MA-d}$.

The last estimator worth discussing in this simulation is the frequentist analog of $\widehat{\Delta}_{DR}^{MA-dII}$. We label this as $\widehat{\Delta}_{DR}^{MS-II}$, which is constructed in the following manner: (1) select the outcome model based on BIC alone; and (2) select the propensity score model from the class of models that excludes covariates that are not included in the chosen outcome model. This estimator performs very similar to that of $\widehat{\Delta}_{DR}^{MA-dII}$ in terms of both bias and MSE. In fact, the two estimators are asymptotically equivalent.

A.2.3 Acknowledgments

Support for this research was provided by National Institute of Environmental Health Sciences grants 5T32ES007142 and R01-ES012054, National Cancer Institute grant P01-CA134294, Environmental Protection Agency grant RD-83479801, and a Health Effects Institute grant (Dominici). We would like to thank Eric Tchetgen Tchetgen for his useful discussions.

A.3 Bias inflation due to exposure prediction in environmental epidemiology

A.3.1 Bias inflation due to measurement error

The main results of our paper rely on the fact that our predicted exposure follows a Berkson error model, and an extension of our results into the case of a Berkson error model is straightforward. Consider the true exposure X is measured with error, and that the measured exposure X^* follows the Berkson error model:

$$X_i = X_i^* + \epsilon_i^*$$

where ϵ_i^* is a mean zero error term that is uncorrelated with X_i^*. Let Y_i be as in Equation 4.1 and consider estimating β_0 using the misspecified regression model $Y_i = \beta X_i^* + \epsilon_i$. The bias of the least squares estimate $\widehat{\beta}^*$ of β is given by:

$$bias(\widehat{\beta}^*) = \mathrm{E}[\widehat{\beta}^* - \beta_0] = bias(\widehat{\beta}_x)\frac{\sigma_x^2}{\sigma_{x^*}^2} \tag{4.1}$$

where $\sigma_{x^*}^2 = var(X^*)$. Note that the expression given in Equation 4.1 is precisely the same as given in Equation 4.3.

This slightly more general result is quite interesting. When there is uncontrolled confounding and an exposure is used that is measured with error (Berkson error), then the bias of the health effect is the product of two pieces: (1) the bias due to lack of adjustment for confounding; and (2) a bias inflation factor that is the ratio of the true variance of the exposure to that of error prone exposure.

Now consider a classical measurement error scenario; the measured exposure is related to the true exposure by

$$X^* = X + \eta$$

where η is a mean zero error term that is uncorrelated with X. Schwartz and Coull (2003) provide a discussion of this issue in the context of controlling for confounding due to multiple exposures, but their results apply if we treat one exposure as confounders. Specifically, it can be shown that the expected value of a health effect estimate when using an exposure that has classical measurement error is given by:

$$\frac{\sigma_x^2}{\sigma_x^2 + \tau^2}(\beta + \xi\gamma) + \frac{1}{\sigma_x^2 + \tau^2} cov(\eta, Y)$$

where $\tau^2 = var(\eta)$ and the p^{th} element of ξ is given by $C_{ip} = \xi_p X_i + \epsilon_i$. Note that using this notation, $bias(\widehat{\beta}_x) = \xi\gamma$. Under the common assumption that the measurement error is non-differential on the outcome, then the expression simplifies to be:

$$\frac{\sigma_x^2}{\sigma_x^2 + \tau^2}(\beta + \xi\gamma)$$

Typically, the term $\frac{\sigma_x^2}{\sigma_x^2+\tau^2}$ is referred to as an attenuation factor, as it attenuates the estimated effect $\mathrm{E}[\widehat{\beta}_x] = \beta + \xi\gamma$ towards zero.

A.3.2 Bias inflation when confounding has been partially controlled or different subsets of confounders are used to predict exposure

In this discussion, we consider four types of covariates: (1) those unrelated to outcome or exposure; (2) those related to outcome but not exposure; (3) those related to exposure but not outcome; and (4) those related to both outcome and exposure. Covariates of type (1) and (2) are not interesting in this setting, while (3) should be used to predict exposure and (4) are the confounders that need to be accounted for in the health effects model.

First, consider the same set up as before, with the exposure-outcome-confounder relationship given by Equation 4.1 and 4.2. Let $\mathbf{C} = (\mathbf{C}^{(1)}, \mathbf{C}^{(2)})$ and $\Sigma_C = \text{var}(\mathbf{C}_i) = \begin{pmatrix} \Sigma_1 & \Sigma_{12} \\ \Sigma_{21} & \Sigma_2 \end{pmatrix}$, where each subset of $\mathbf{C}$ could contain any type of covariate. Further, let $W = \mathbf{C}\alpha$ be the predicted exposure if the exposure model from Equation 4.1 were known exactly, $W_1 = \mathbf{C}^{(1)}\alpha_1^*$ be the predicted exposure if the misspecified exposure model $X_i = \mathbf{C}_i^{(1)}\alpha_1^* + \epsilon$ were known exactly, and $W_2 = \mathbf{C}_i^{(2)}\alpha_2^*$ be the predicted exposure if the misspecified exposure model $X_i = \mathbf{C}_i^{(2)}\alpha_2^* + \epsilon$ were known exactly.

Table A.6 provides the bias of the health effect estimate for each choice of the predicted exposure and an outcome model that either fails to control for any confounding ($Y = \beta W + \epsilon$) or an outcome model that controls for only $\mathbf{C}^{(1)}$ ($Y = \beta W + \mathbf{C}^{(1)}\gamma + \epsilon$). Further, let $\widetilde{R}_z^2$ denote the population value of the R^2 from the exposure model that uses arbitrary Z as a prediction of X. Table A.7 provides the R^2 and its corresponding population value for each of the predicted exposures W, W_1, or W_2.

The bias of $\widehat{\beta}_w$ given in Table A.6 is the bias of the health effect estimate provided in Equation 4.3 that was previously described under the situation that the predicted exposure W is used in an outcome model that fails to control for any confounding. Recall that is was shown that this bias is the product of the bias due to lack of adjustment for any confounding and a bias inflation factor due to exposure prediction that is the inverse of the $\widetilde{R}_w^2$.

This relationship holds true for any collection of covariates, regardless of their association with the exposure and the outcome. For example, suppose all $\mathbf{C}$ are only related to the exposure. Then, there is no confounding and as a result, the bias of $\widehat{\beta}_w$ is 0. Similarly, suppose that all $\mathbf{C}$ are only related to the outcome. Then, $\widetilde{R}_w^2 = 0$ because $\mathbf{C}$ has no power to predict exposure, and the bias of $\widehat{\beta}_w$ increases in magnitude to infinity.

Next, consider a situation where the true set of confounders $\mathbf{C}$ is unknown to the researcher but the true exposure X is observed, and instead of controlling for the full set of $\mathbf{C}$s, the decision is made to only control for the subset $\mathbf{C}^{(1)}$ (first row, second column). The

Table A.6: Bias of a health effect estimate when confounding has been partially controlled or different subsets of confounders are used to predict exposure

	Outcome model	
Exposure	$Y = W + \epsilon$	$Y = W + C^{(1)} + \epsilon$
$W = X$	$\mathrm{E}[\widehat{\beta}_x - \beta_0] = b_x = \frac{\alpha^T \Sigma_c \gamma}{\sigma^2_{x\|c} + \alpha^T \Sigma_C \alpha}$	$\mathrm{E}[\widehat{\beta}_x^{(1)} - \beta_0] = b_x^{(1)} = \frac{\alpha_2^T (\Sigma_2 - \Sigma_{21}\Sigma_1^{-1}\Sigma_{12})\gamma_2}{\sigma^2_{x\|c} + \alpha_2^T (\Sigma_2 - \Sigma_{21}\Sigma_1^{-1}\Sigma_{12})\alpha_2}$
$W = C^{(1)}\alpha_1^*$	$\mathrm{E}[\widehat{\beta}_{w_1} - \beta_0] = b_{w_1} = b_x^{(1)} + (b_x - b_x^{(1)})\widetilde{R}^{-2}_{w_1}$	NA[a]
$W = C^{(2)}\alpha_2^*$	$\mathrm{E}[\widehat{\beta}_{w_2} - \beta_0] = b_{w_2} = b_x^{(2)} + (b_x - b_x^{(2)})\widetilde{R}^{-2}_{w_2}$	$\mathrm{E}[\widehat{\beta}_{w_2}^{(1)} - \beta_0] = \beta_0 \frac{\alpha_2^{*T}\Sigma_{21}\Sigma_1^{-1}\Sigma_{12}\alpha_2^* - \alpha_1^{*T}\Sigma_{12}\alpha_2^*}{\alpha_2^{*T}(\Sigma_2 - \Sigma_{21}\Sigma_1^{-1}\Sigma_{12})\alpha_2^*} + \frac{\alpha_2^{*T}(\Sigma_2 - \Sigma_{21}\Sigma_1^{-1}\Sigma_{12})\gamma_2}{\alpha_2^{*T}(\Sigma_2 - \Sigma_{21}\Sigma_1^{-1}\Sigma_{12})\alpha_2^*}$
$W = C\alpha$	$\mathrm{E}[\widehat{\beta}_w - \beta] = b_w = b_x \widetilde{R}^{-2}_w$	NA[a]

[a] Does not exist due to collinearity between predicted exposure and confounding adjustment.

Table A.7: Coefficient of determination (R^2) and its corresponding population value ($\widetilde{R}^2$)

Predicted exposure	**Coefficient of determination**	**Population based** R^2
$W = \mathbf{C}\alpha_0$	$R^2_w = \frac{\sum(W_i - X_i)^2}{\sum(X_i - \bar{X})^2}$	$\widetilde{R}^2_w = \frac{\alpha^T \Sigma_C \alpha}{\sigma^2_{x\|c} + \alpha^T \Sigma_c \alpha} = \frac{\sigma^2_w}{\sigma^2_x}$
$W_1 = \mathbf{C}^{(1)}\alpha^*_1$	$R^2_{w_1} = \frac{\sum(W_{1i} - X_i)^2}{\sum(X_i - \bar{X})^2}$	$\widetilde{R}^2_{w_1} = \frac{\alpha^{*T}_1 \Sigma_1 \alpha^*_1}{\sigma^2_{x\|c} + \alpha^T \Sigma_c \alpha} = \frac{\sigma^2_{w_1}}{\sigma^2_x}$
$W_2 = \mathbf{C}^{(2)}\alpha^*_2$	$R^2_{w_2} = \frac{\sum(W_{2i} - X_i)^2}{\sum(X_i - \bar{X})^2}$	$\widetilde{R}^2_{w_2} = \frac{\alpha^T \Sigma_C \alpha}{\sigma^2_{x\|c} + \alpha^T \Sigma_c \alpha} = \frac{\sigma^2_{w_1}}{\sigma^2_x}$

bias of the health effect estimate from the misspecified outcome model $Y = \beta X + \gamma \mathbf{C}^{(1)} + \epsilon$ is given by $bias(\widehat{\beta}_x^{(1)})$ in Table A.6. This corresponds to the bias due to the failure to control for the confounding due to $\mathbf{C}^{(2)}$. In other words, it is the bias due to confounding that remains after controlling for $\mathbf{C}^{(1)}$, but failing to control for the full set of necessary confounders $\mathbf{C}$. Suppose that $\mathbf{C}^{(1)}$ contains all covariates that are confounders and $\mathbf{C}^{(2)}$ contains any remaining covariates. Then, $bias(\widehat{\beta}_x^{(1)}) = 0$ because confounding has sufficiently been controlled by $\mathbf{C}^{(1)}$ alone. However, suppose that $\mathbf{C}^{(2)}$ contains all covariates that are confounders, $\mathbf{C}^{(1)}$ contains any remaining covariates, and $\mathbf{C}^{(1)}$ and $\mathbf{C}^{(2)}$ are uncorrelated. Then, $bias(\widehat{\beta}_x^{(1)}) = bias(\widehat{\beta}_x)\widetilde{R}_{w_2}^{-2}$ so that the bias of the health effect estimate is inflated by controlling for covariates that are not confounders. This is a specific example of bias inflation that arises from conditioning on instrumental variables.[1,2]

Now consider a situation where the true exposure X is unobserved, and instead is predicted with a subset of the Cs (second row, first column). The $bias(\widehat{\beta}_{w_1})$ is the bias of the health effect estimate in the situation that the predicted exposure $W_1 = \mathbf{C}^{(1)}\alpha_1^*$ is used in the outcome model that fails to control for any confounding. From Table A.6, we note that this bias decomposes into two parts, with the first one being the bias due to the failure to control for confounding due to $\mathbf{C}^{(2)}$. Therefore, ignoring the second term, using $\mathbf{C}^{(1)}$ to predict the exposure appears to help control the confounding due to $\mathbf{C}^{(1)}$. However, this is not exactly the case, as the second term of $bias(\widehat{\beta}_{w_1})$ in Table A.6 can either decrease or increase the magnitude of the bias. Further we note that $bias(\widehat{\beta}_{w_1})$ depends on the inverse of $\widetilde{R}_{w_1}^2$; therefore, the bias of $\widehat{\beta}_{w_1}$ is a function of how well W_1 predicts X. As $\widetilde{R}_{w_1}^2$ goes to 1, $bias(\widehat{\beta}_{w_1}) = bias(\widehat{\beta}_x)$, so that if W_1 predicts X perfectly, we are left with the bias due to lack of adjustment for confounding in the situation where the true exposure X is known. Similarly, as $\widetilde{R}_{w_1}^2$ goes to 0, the $bias(\widehat{\beta}_{w_1})$ increases in magnitude to infinity, suggesting that if we cannot accurately predict the exposure, we cannot return a valid effect estimate. However, as $\widetilde{R}_{w_1}^2$ varies between 0 and 1, no general statement can be made about the magnitude of the bias. Similar results hold for $bias(\widehat{\beta}_{w_2})$.

Suppose that $\mathbf{C}^{(1)}$ contains all covariates that are confounders and $\mathbf{C}^{(2)}$ contains any re-

maining covariates. Then, $bias(\widehat{\beta}_{w_1}) = bias(\widehat{\beta}_x)\widetilde{R}^{-2}_{w_1}$, or in other words, we have an expression similar to $bias(\widehat{\beta}_w)$ in that we are inflating the bias due to lack of adjustment for confounding. By moving covariates that are not confounders from $\mathbf{C}^{(2)}$ into $\mathbf{C}^{(1)}$, we would increase $\widetilde{R}^2_{w_1}$ and as a result $bias(\widehat{\beta}_{w_1})$ would decrease. Therefore, if all confounders are used to predict the exposure, we decrease the bias of the health effect estimate by improving the prediction accuracy.

The last situation provided in Table A.6 is a situation where the true exposure X is unobserved, instead is predicted with a subset of the Cs, and a different set of Cs are used to control confounding in the outcome model (third row, second column). Specifically, the $bias(\widehat{\beta}^{(1)}_{w_2})$ is the bias of the health effect estimate in the situation that the predicted exposure $W_2 = \mathbf{C}^{(2)}\alpha^*_2$ is used in the outcome model that controls for only $\mathbf{C}^{(1)}$ ($Y = \beta W_2 + \mathbf{C}^{(1)}\gamma + \epsilon$). We wish to only point out a few features of the expression for this bias. First, the bias depends on the true underlying effect β_0. As the true effect size increases, so does the magnitude of bias. Second, the expression for the bias of $\widehat{\beta}^{(1)}_{w_2}$ is much more complex than any of the other biases given in Table A.6 and will not be described in detail. However, suppose again that $\mathbf{C}^{(1)}$ contains all covariates that are confounders and $\mathbf{C}^{(2)}$ contains any remaining covariates. Further, assume that $\mathbf{C}^{(1)}$ and $\mathbf{C}^{(2)}$ are uncorrelated. Then, $bias(\widehat{\beta}^{(1)}_{w_2}) = 0$. This occurs because: (1) confounding has been sufficiently controlled through $\mathbf{C}^{(1)}$; and (2) the exposure is predicted with covariates that are uncorrelated with confounders. However, if $\mathbf{C}^{(1)}$ and $\mathbf{C}^{(2)}$ are correlated, then $bias(\widehat{\beta}^{(1)}_{w_2}) \neq 0$.

Considering these results, if we can separate our covariates into two orthogonal sets, one of which contains all necessary confounders, then we can hope to construct an exposure prediction model along with an outcome regression model that yield an unbiased health effect estimate.

The biases given in Table A.6 are difficult to compare, except for in the simplest situations as in $bias(\widehat{\beta}_x)$ and $bias(\widehat{\beta}_w)$. Therefore, it is difficult to make any general conclusions about whether including or excluding a potential confounder from either the exposure model

or the outcome model is beneficial or detrimental to the final goal of effect estimation.

The previous point warrant further discussion; when the goal of a study is effect estimation, the decision to include or exclude a potential confounder from either the outcome or the exposure model needs to be based on more than just the predictive power of the potential confounder on the exposure or the strength of the relationship with the outcome, but instead the decision needs to be based on some tradeoff between the two. Current statistical methods for model selection fail in this regard, as they have been designed to control confounding and ignore exposure prediction all together.

A.3.3 Additional simulations

Following the simulation setup of the Section 4.3 exactly, we provide additional simulated results for two additional choices of the parameter γ. Specifically, let

$$\gamma^a = (0, -0.044, -0.075, 0.105, 0.090, -0.082, 0.096, 0.0897, -0.041, 0.011)$$
$$\gamma^b = (0.025, 0.0067, -0.0058, 0.005, 0.0208, 0.0033, 0.025, 0.025, 0.0125, 0)$$

The purpose of these two additional specifications is to illustrate that in some cases, increasing the R^2 always decreases the bias, while in others, increasing the R^2 always increases the bias. From Figure A.1, we note that the bias increases with the R^2. Therefore, adding additional covariates to the exposure prediction model adds bias to the estimated health effect. From Figure A.2, we note that the bias decreases as R^2 increases. Therefore, adding additional covariates to the exposure prediction model improves the health effect estimate.

These results, in addition to those in the main text, provide evidence that bias inflation due to exposure prediction can either reduce or increase the bias of the health effect estimate. Therefore, it is not possible to make general conclusions as to whether better

exposure prediction models will lead to better health effect estimates.

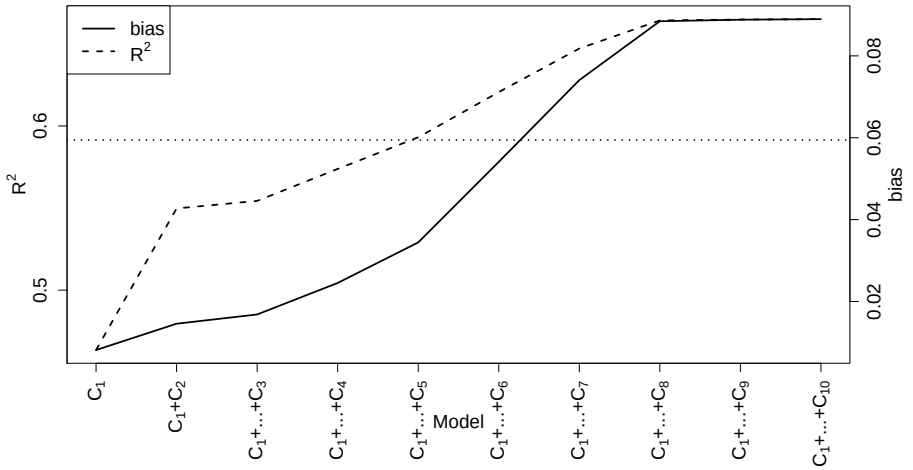

Figure A.1: Tradeoff between R^2 and bias from the hypothetical cohort study of the association between long-term exposure to $PM_{2.5}$ and cardiovascular hospitalization rates in the New England region under γ^a

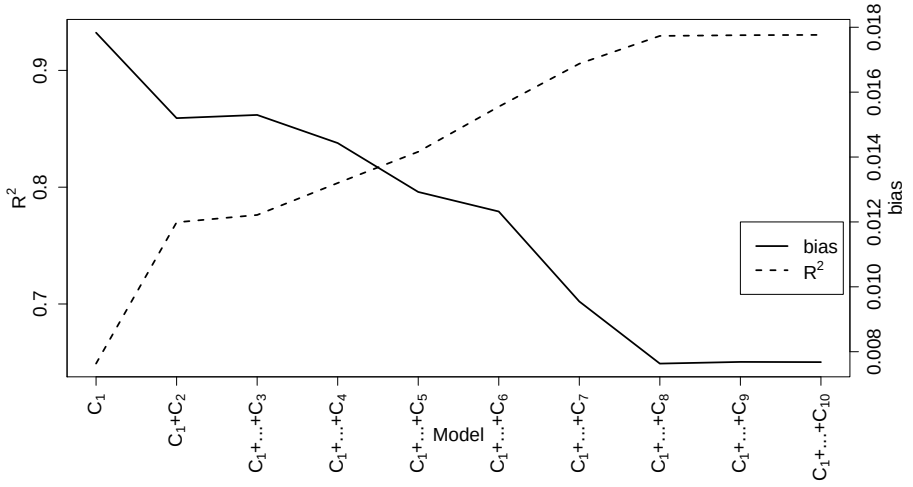

Figure A.2: Tradeoff between R^2 and bias from the hypothetical cohort study of the association between long-term exposure to $PM_{2.5}$ and cardiovascular hospitalization rates in the New England region under γ^b

A.3.4 Acknowledgements

We would like to thank Itai Kloog for providing the data for the simulation, and Arden Pope and Jennifer Bobb for the useful feedback. Support for this research was provided by NIH grants T32ES007142, R21 ES021427, R01 ES019955, and R21 ES020152, EPA grants RD 83490001, RD 83479801, and R834894, NCI grant P01 CA134294-02, and HEI grant 4909.

References

Bang, H. and J. M. Robins (2005). "Doubly robust estimation in missing data and causal inference models," *Biometrics*, 61, 962–973.

Bhattacharya, J. and W. B. Vogt (2012). "Do instrumental variables belong in propensity scores?" *International Journal of Statistics & Economics*, 9, 107–127.

Bickel, P., C. Klaassen, Y. Ritov, and J. Wellner (1998). *Efficient and Adaptive Estimation for Semiparametric Models*, Johns Hopkins series in the mathematical sciences, Springer.

Breysse, P., R. Delfino, F. Dominici, A. Elder, M. Frampton, J. Froines, A. Geyh, J. Godleski, D. Gold, P. Hopke, P. Koutrakis, G. Li, Ning Oberdrster, K. Pinkerton, J. Samet, M. Utell, and A. Wexler (2012). "Us epa particulate matter research centers: summary of research results for 20052011," *Air Qual Atmos Health*.

Brookhart, M. A. (2006). "Variable Selection for Propensity Score Models," *American Journal of Epidemiology*, 163, 1149–1156.

Cao, W., A. A. Tsiatis, and M. Davidian (2009). "Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data," *Biometrika*, 96, 723–734.

Carroll, R. J., D. Ruppert, and L. A. Stefanski (1995). *Measurement Error in Nonlinear Models*, Chapman and Hall/CRC.

Cesaroni, G., C. Badaloni, C. Gariazzo, M. Stafoggia, R. Sozzi, M. Davoli, F. Forastiere, et al. (2013). "Long-term exposure to urban air pollution and mortality in a cohort of more than a million adults in rome," *Environmental health perspectives*.

Chu, H. and S. Cole (2010). "Estimation of risk ratios in cohort studies with common outcomes. a bayesian approach," *Epidemiology*, 21, 855–862.

Crainiceanu, C. M., F. Dominici, and G. Parmigiani (2008). "Adjustment uncertainty in effect estimation," *Biometrika*, 95, 635–651.

Dominici, F., A. McDermott, and T. J. Hastie (2004). "Improved semiparametric time series models of air pollution and mortality," *Journal of the American Statistical Association*, 99, 938–948.

Dominici, F., L. Sheppard, and M. Clyde (2003). "Health effects of air pollution: A statistical review," *International Statistical Review*, 71, 243–276.

Drake, C. (1993). "Effects of misspecification of the propensity score on estimators of treatment effect," *Biometrics*.

Draper, D. (1995). "Assessment and propagation of model uncertainty," *Journal of the Royal Statistical Society. Series B (Methodological)*, 57, pp. 45–97.

Efron, B. and R. Tibshirani (1993). *An Introduction to the Bootstrap*, New York, NY: Chapman & Hall.

Friedman, J., T. Hastie, and R. Tibshirani (2008). *The elements of statistical learning*, volume 2, Springer.

George, E. and R. McCulloch (1993). "Variable selection via gibbs sampling," *Journal of the American Statistical Association*, 88, 881–889.

Greenland, S. (2004). "Model-based estimation of relative risks and other epidemiologic measures in studies of common outcomes and in case-control studies," *American Journal of Epidemiology*, 160, 301–305.

Gryparis, A., C. J. Paciorek, A. Zeka, J. Schwartz, and B. A. Coull (2009). "Measurement error caused by spatial misalignment in environmental epidemiology," *Biostatistics*, 10, 258–274.

Henderson, S. B., B. Beckerman, M. Jerrett, and M. Brauer (2007). "Application of land use regression to estimate long-term concentrations of traffic-related nitrogen oxides and fine particulate matter," *Environmental science & technology*, 41, 2422–2428.

Hirano, K., G. Imbens, and G. Ridder (2003). "Efficient estimation of average treatment effects using the estimated propensity score," *Econometrica*, 71, 1161–1189.

Hoek, G., R. Beelen, K. de Hoogh, D. Vienneau, J. Gulliver, P. Fischer, and D. Briggs (2008). "A review of land-use regression models to assess spatial variation of outdoor air pollution," *Atmospheric environment*, 42, 7561–7578.

Hoeting, J. A., D. Madigan, A. E. Raftery, and C. T. Volinsky (1999). "Bayesian model averaging: a tutorial (with comments by m. clyde, david draper and e. i. george, and a rejoinder by the authors)," *Statistical Science*, 14, 382–417.

Johnson, V. and D. Rossell (2012). "Bayesian model selection in high-dimensional settings," *Journal of the American Statistical Association*, 107, 649–660.

Kamo, N., M. Carlson, R. Brennan, and F. Earls (2008). "Young citizens as health agents: Use of drama in promoting community efficacy for hiv/aids," *Journal Information*, 98.

Kass, R. and A. Raftery (1995). "Bayes factors," *Journal of the american statistical association*, 90, 773–795.

Kloog, I., B. A. Coull, A. Zanobetti, P. Koutrakis, and J. D. Schwartz (2012a). "Acute and chronic effects of particles on hospital admissions in new-england," *PloS one*, 7, e34664.

Kloog, I., S. Melly, W. Ridgway, B. Coull, and J. Schwartz (2012b). "Using new satellite based exposure methods to study the association between pregnancy pm2.5 exposure, premature birth and birth weight in massachusetts," *Environmental Health*, 11, 40.

Knol, M. J., R. G. Duijnhoven, D. E. Grobbee, K. G. Moons, and R. H. Groenwold (2011). "Potential misinterpretation of treatment effects due to use of odds ratios and logistic regression in randomized controlled trials," *PLoS One*, 6, e21248.

Konishi, S. and G. Kitagawa (1996). "Generalised information criteria in model selection," *Biometrika*, 83, 875–890.

Lee, J. (1994). "Odds ratio or relative risk for cross-sectional data?" *International Journal of Epidemiology*, 23, 201–203.

Liang, K.-Y. and S. L. Zeger (1986). "Longitudinal data analysis using generalized linear models," *Biometrika*, 73, 13–22.

Lunceford, J. and M. Davidian (2004). "Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study," *Statistics in Medicine*, 23, 2937–2960.

Madsen, L., D. Ruppert, and N. Altman (2008). "Regression with spatially misaligned data," *Environmetrics*, 19, 453–467.

Neupane, B., M. Jerrett, R. T. Burnett, T. Marrie, A. Arain, and M. Loeb (2010). "Long-term exposure to ambient air pollution and risk of hospitalization with community-acquired pneumonia in older adults," *American journal of respiratory and critical care medicine*, 181, 47–53.

Nishii, R. (1984). "Asymptotic properties of criteria for selection of variables in multiple regression," *The Annals of Statistics*, 12, 758–765.

O'Hara, R. and M. Sillanpää (2009). "A review of bayesian variable selection methods: what, how and which," *Bayesian Analysis*, 4, 85–117.

Oliver, M. A. and R. Webster (1990). "Kriging: a method of interpolation for geographical information systems," *International Journal of Geographical Information System*, 4, 313–332.

Pearl, J. (2012). "On a class of bias-amplifying variables that endanger effect estimates," *arXiv preprint arXiv:1203.3503*.

Pope, C. A. (2007). "Mortality effects of longer term exposures to fine particulate air pollution: Review of recent epidemiological evidence," *Inhalation Toxicology*, 19, 33–38.

Pope III, C. A. and R. T. Burnett (2007). "Confounding in air pollution epidemiology: the broader context," *Epidemiology*, 18, 424–426.

Raftery, A., D. Madigan, and J. Hoeting (1997). "Bayesian model averaging for linear regression models," *Journal of the American Statistical Association*, 92, 179–191.

Robins, J., M. Hernan, and B. Brumback (2000). "Marginal structural models and causal inference in epidemiology," *Epidemiology*, 11(5), 550–560.

Robins, J., M. Sued, Q. Lei-Gomez, and A. Rotnitzky (2007). "Comment: Performance of double-robust estimators when"inverse probability" weights are highly variable," *Statistical Science*, 22, 544–559.

Rosenbaum, P. R. and D. B. Rubin (1983). "The central role of the propensity score in observational studies for causal effects," *Biometrika*, 70, 41–55.

Rosenbaum, P. R. and D. B. Rubin (1984). "Reducing bias in observational studies using subclassification on the propensity score," *Journal of the American Statistical Association*, 79, 516–524.

Ross, Z., M. Jerrett, K. Ito, B. Tempalski, and G. D. Thurston (2007). "A land use regression for predicting fine particulate matter concentrations in the new york city region," *Atmospheric Environment*, 41, 2255–2269.

Rubin, D. et al. (1997). "Estimating causal effects from large data sets using propensity scores," *Annals of internal medicine*, 127, 757–763.

Rubin, D. B. (1974). "Estimating causal effects of treatments in randomized and nonrandomized studies," *Journal of educational Psychology*, 66, 688–701.

Sahsuvaroglu, T., M. Jerrett, M. R. Sears, R. McConnell, N. Finkelstein, A. Arain, B. Newbold, R. Burnett, et al. (2009). "Spatial analysis of air pollution and childhood asthma in hamilton, canada: comparing exposure methods in sensitive subgroups," *Environmental Health*, 8, 14.

Scharfstein, D. O., A. Rotnitzky, and J. M. Robins (1999). "Reply to comments on "Adjusting for nonignorable drop-out using semiparametric nonresponse models"," *Journal of the American Statistical Association*, 94, 1135–1146.

Schneeweiss, S., J. A. Rassen, R. J. Glynn, J. Avorn, H. Mogun, and M. A. Brookhart (2009). "High-dimensional propensity score adjustment in studies of treatment effects using health care claims data," *Epidemiology*, 20, 512–22.

Schwartz, J. and B. A. Coull (2003). "Control for confounding in the presence of measurement error in hierarchical models," *Biostatistics*, 4, 539–553.

Schwarz, G. (1978). "Estimating the Dimension of a Model," *The Annals of Statistics*, 6, 461–464.

Sheppard, L., R. Burnett, A. Szpiro, S. Kim, M. Jerrett, C. Pope, and B. Brunekreef (2011). "Confounding and exposure measurement error in air pollution epidemiology," *Air Quality, Atmosphere & Health*, 1–14.

Skove, T., J. Deddens, M. R. Petersen, and L. Endahl (1998). "Prevalence proportion ratios: estimation and hypothesis testing," *International Journal of Epidemiology*, 27, 91–95.

Smith, A. and D. Spiegelhalter (1980). "Bayes factors and choice criteria for linear models," *RoyalStat*, B-42, 213–220.

Spiegelman, D. and E. Hertzmark (2005). "Easy sas calculations for risk or prevalence ratios and differences," *American Journal of Epidemiology*, 162, 199–200.

Stephens, A., E. Tchetgen Tchetgen, and V. Gruttola (2012). "Augmented generalized estimating equations for improving efficiency and validity of estimation in cluster randomized trials by leveraging cluster-level and individual-level covariates," *Statistics in Medicine*.

Szpiro, A., C. Paciorek, and L. Sheppard (2011a). "Does more accurate exposure prediction necessarily improve health effect estimates?" *Epidemiology*, 22, 680–685.

Szpiro, A. A., L. Sheppard, and T. Lumley (2011b). "Efficient measurement error correction with spatially misaligned data," *Biostatistics*, 12, 610–623.

Tan, Z. (2010). "Bounded, efficient and doubly robust estimation with inverse weighting," *Biometrika*, 97, 661–682.

Tchetgen Tchetgen, E. (2012). "Estimation of risk ratios in cohort studies with common outcomes: A simple and efficient two-stage approach," *International Journal of Biostatistics*, In press.

van der Laan, M. (2010). "Targeted maximum likelihood based causal inference: part i," *The International Journal of Biostatistics*, 6.

van der Laan, M., E. Polley, and A. Hubbard (2007). "Super learner," *Statistical Applications in Genetics and Molecular Biology*, 6.

Vansteelandt, S., M. Bekaert, and G. Claeskens (2010). "On model selection and model misspecification in causal inference." *Statistical methods in medical research*.

Wacholder, S. (1986). "Binomial regression in glim: Estimating risk ratios and risk differences," *American Journal of Epidemiology*, 123, 174–184.

Wang, C., G. Parmigiani, and F. Dominici (2012a). "Bayesian effect estimation accounting for adjustment uncertainty," *Biometrics*.

Wang, C., G. Parmigiani, and F. Dominici (2012b). "Rejoinder: Bayesian effect estimation accounting for adjustment uncertainty," *Biometrics*.

Wasserman, L. (2005). *All of nonparametric statistics*, Springer.

Yanosky, J. D., C. J. Paciorek, J. Schwartz, F. Laden, R. Puett, and H. H. Suh (2008). "Spatio-temporal modeling of chronic pm10 exposure for the nurses health study," *Atmospheric Environment*, 42, 4047 – 4062.

Yelland, L. N., A. B. Salter, and P. Ryan (2011). "Performance of the modified poisson regression approach for estimating relative risks from clustered prospective data," *American Journal of Epidemiology*, 174, 984–992.

Zigler, C. and F. Dominici (2012). "Uncertainty in propensity score estimation: Bayesian methods for variable selection and model averaged causal effects," *Technical Report*.

Zou, G. (2004). "A modified poisson regression approach to prospective studies with binary data," *American Journal of Epidemiology*, 159, 702–706.